Universidade de Aveiro Departamento de Electrónica,
2014 Telecomunicações e Informática

André Pereira de Matos

Identificação de conceitos de saúde em redes sociais

Identification of health concepts in social networks

Universidade de Aveiro Departamento de Electrónica
2014 Telecomunicações e Informática

André Pereira de Matos

Identificação de conceitos de saúde em redes sociais

Identification of health concepts in social networks

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Professor Doutor José Luís Oliveira, Professor Associado do Departamento de Electrónica Telecomunicações e Informática da Universidade de Aveiro, e do Doutor Sérgio Aleixo Matos, Investigador Auxiliar do Instituto de Engenharia Electrónica e Telemática de Aveiro.

Dedico este trabalho aos meus pais, irmã e avós que ainda me vão acompanhando nestas andanças. Nunca esquecendo os amigos que não param de me chatear e os colega que estão sempre prontos a dar uma ajudinha se for caso disso.

This work is dedicated to my parents, sister and grandparents who are still present in another milestone of my life. To the friends who don't go unnoticed and still annoy me every now and again, and to all colleagues who are always ready to lend a hand if the situation requires.

o júri

presidente / president

Prof. Dr. Augusto Silva
Professor Auxiliar do Departamento de Electrónica Telecomunicações e Informática da Universidade de Aveiro

vogais / examiners committee

Prof. Dr. Joel P. Arrais
Professor Auxiliar Convidado do Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Dr. Sérgio Aleixo Matos
Investigador Auxiliar do Instituto de Engenharia Electrónica e Telemática de Aveiro

agradecimentos

Em primeiro lugar, agradeço toda a ajuda aos meus orientadores José Luís Oliveira e Sérgio Matos e também ao David Campos por, ao início, ter sido o responsável por me ajudar a integrar no grupo de Bioinformática.

À minha família, mais especificamente Pais, Irmã e Avós, obrigado por tudo e mais alguma coisa!

Aos Leirienses, minha família emprestada, já não era sem tempo, hein?

palavras-chave

mineração de dados, redes sociais, medicamentos, doenças, patologias, químicos.

resumo

No últimos anos, blogs e redes sociais tiveram um crescimento imensurável, tanto em número de utilizadores como em receita gerada. Por outro lado, portais como o Google, Yahoo e Bing tornaram-se interfaces privilegiadas de procura de informação a nível global. Sob um ponto de vista empresarial, este fenómeno tem igualmente levado à mudança de estratégias de representação, indexação e pesquisa de informação. Surgiram motores como Lucene, Nutch, Solr, Sphinx, ElasticSearch, entre muitos outros, que são hoje sistemas incontornáveis na engenharia do software. A informação que diariamente é disponibilizada na Internet pode ter um enorme valor em múltiplas vertentes - social, científica, política, económica. Por exemplo, na área da saúde, as instituições registam cada vez mais informação sobre utentes, procedimentos, exames, diagnósticos, etc. Instituições governamentais divulgam cada vez mais informação sobre saúde nos seus mais variados aspectos. Por outro lado, a investigação em saúde é uma das áreas mais ativas, resultando na publicação em rede de novos resultados. Para além disto, a forma como os cidadãos trocam informação sobre o seu estado de saúde pode igualmente ajudar a detectar pandemias bem como a identificar condições clínicas e as suas causas.

Neste projeto pretendeu-se desenvolver um sistema capaz de processar informação proveniente de redes sociais, identificando sinais, sintomas, doenças e medicamentos. Do ponto de vista de engenharia, o processo consistiu em aceder a diversas fontes de dados para obter informação, criar dicionários e modelos de processamento de linguagem natural para identificar padrões no texto e associações entre conceitos. Associações essas que, futuramente, poderão ser um passo para a detecção de reacções adversas a medicamentos.

keywords data mining, social networks, drugs, diseases, conditions, chemicals.

abstract

Social Networks are, without doubt, one of the fastest growing trends on the Internet nowadays and the amount of information generated is huge. On the other hand, search engines such as Google, Yahoo and Bing became global interfaces for information search. From an enterprise point of view, this phenomenon has lead to changes in strategies to deal with information representation, indexing and searching. Engines like Lucene, Nutch, Solr, Sphinx, ElasticSearch, among others, were born and became important platforms of software engineering. The amount of information available on the Internet on a daily basis, can be utterly valuable for multiple goals - social, scientific, politic, economic. For instance, in the case of health related subjects, institutions keep information on patients such as pathologies, conditions, treatments, exam results, pathways, etc. Also, governmental institutions divulge more and more information on health subjects in the most varied aspects. On the other hand, health research is one of the most active areas, resulting on network publishing of new results. Adding to the aforementioned, the way people share information on their health status can be valuable to detect pandemics and also to identify clinical conditions and their causes.

In this project we proposed to develop a system capable of processing information retrieved from social networks, identifying symptoms, disorders and drugs. From an engineering point of view, the process consisted in accessing varied data sources to obtain information, create dictionaries and models of natural language processing to identify text patterns and associations between concepts. In the future, these associations can be valuable, for instance, to detect adverse reaction to drugs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social networks are, without any doubt, one of the fastest growing trends on the Internet nowadays and the amount of information generated is huge. Using that information for identifying trends and exploring its meaning for certain goals can be useful in a variety of contexts. This information can be useful for either social, scientific, political or even economical fields and, in our case, for health related subjects. In the particular case of health related subjects, since Hospitals and other health institutions keep information from patient pathologies, conditions, exam results, pathways, there is a growing knowledge on the most varied health contexts. This information is utterly valuable for studies on pathologies, drugs and their effects, ways to improve treatments for a given condition or disease, and much more. This is why health research programs are becoming more and more common: their potential cannot be ignored and as the most powerful health related entities become aware of such potential, its importance starts to spread and a number of data mining systems start to appear from the most varied institutions. Data mining systems can be used on a multitude of different sources such as the aforementioned social networks, newspapers, forums, newsletters among others and also for very different contexts. Regarding sources, each has their strengths and their limitations and it is up to the researcher to find, discuss and choose which are the ones more prone to the desired goal. In this work, we propose an architecture for a system which is capable of acquiring data from different sources, analyze the information retrieved and be able to find co-relations between health event concepts such as drugs, disorders and symptoms.

Chapter 2

Data Mining in Public Health

In the last few years, the Internet has grown in such ways that one cannot ignore its potential as source of information on a variety of contexts such as politics, economics, scientific research, health events. In each of these areas, we have an infinite number of subtopics to which we can address our goals, for instance, regarding health events one can try to infer flu trends in specific geographic areas, or try to understand which drugs are used for this or that disease/symptom; regarding politics, one can try to learn which set of people is more prone to vote for this or that candidate on some election.

In this work, the main focus will be on health events: their importance, problems and some technical analysis.

2.1 Importance of Data Mining in Public Health

In recent years, due to the growth of Information Systems used to store information about patients and their medical background, we have achieved an enormous quantity of information that can be used to prevent diseases and/or improve the knowledge on a variety of subjects. According to some specialists, the usage of data mining techniques on the available data can prevent hospital errors and achieve new life-saving information [12]. From an economic point of view, if more information can be retrieved from data that already exists, more money can be saved and, hopefully, that money can be used for other purposes. Other area where data mining is relevant has to do with early detection/prevention of diseases: some symptoms can be related to some form of degeneration leading to some specific disease like heart problems, kidney failures or depression. Still regarding the prevention area, data mining can be used to prevent pandemic diseases by detecting risk factors well before they would be noticeable

to health authorities or, in the worst case scenario, to provide a quicker and more effective response in the event of an epidemic outburst.

On the other hand, one of the biggest setbacks of data mining usage for public health has to do with the source of the available data and how that data is to be used. Data should preserve all privacy and confidentiality as to patients' related information is concerned; failing to do so, would become a major legal problem. Regarding studies and conclusions, all results obtained through data mining shall be very well overseen by a specialist to ensure that those results are really an improvement and do not propose risks for future applications.

2.1.1 World Wide Web and Disease Monitoring

Disease monitoring and detection is one valuable tool for any country or global health organization and has been used since the first forms of health related internet-based systems started to appear. ProMED (Program to Monitor Emerging Diseases) was started in 1993 with the aim of building a system to be physically based on different countries in form of strategically-located institutions whose goal was to monitor endemic and emerging diseases and to act as a sentinel so to warn other peers using network based conferences. In 1994, with over 5000 participants in over 110 countries, ProMED-mail, once a trial project, became a permanent and independent reporting system [54]. Although this system relies on reports by various expert teams manually monitoring, reviewing and investigating diverse sources of information such as media reports, official reports, locally observed phenomenons, and thus, not being an automatic system, its pioneer status deserved its mention in this work. Nowadays ProMED-mail has around 60000 subscribers and is active in 185 countries.

Canada also has its own monitoring system: "Global Public Health Intelligence Network" (GPHIN), developed by the Public Health Agency of Canada. GPHIN gathers reports of public health from news media networks, news wires, newspapers, etc, on a real-time basis and in 8 different languages. The information is filtered for relevance using an automated process and, after that, is complemented by human analysis. The relevant output is then categorized and made available to relevant health authorities [50].

Other sensible approach to this thematic is using the powerful Google and Yahoo search engine trends. Eysenbach proved to exist an excellent correlation between the number of clicks on a keyword-triggered link in Google with epidemiological data from 2004/2005 in Canada. Mainly tracking flu-related searches using the words "flu" or "flu symptoms" and using Google statistics to determine the number of hits of each search string and comparing with "influenza like illnesses" report numbers published by the Public Health Agency Canada, the author

concluded that results provided a promising method for future and more refined studies for early warning systems on infectious disease outbreaks, bio-terrorism or emerging diseases [33]. A similar approach was followed by Polgreen et al. in a study relating the searches for influenza using yahoo and the physical influenza occurrence. By counting the weekly queries originating from only the United States and containing influenza-related search terms, linear models were estimated. In the end, only by using the frequency of searches, the system predicted an increase in cultures positive for influenza 1 to 3 weeks in advance of when they actually occurred [63].

There are other systems relying on different sources of information like, for instance, the Global Health Monitor by Doan et al. [28]. This system analyzes English news stories from news feed providers, classifies them for topic relevance and plots them into a Google map allowing public health workers to monitor the spread of diseases in a geographic and temporal context.

Although not entirely related to disease monitoring but being important in the sense of monitoring the spread of a given treatment, a study by Nakada et al. explained why the human papillomavirus (HPV) vaccination was achieved relatively quickly in Japan as compared to United States and India [56]. By selecting keywords on Japanese newspapers and web pages during a period of time, and analyzing their relevance, it was concluded that the "rapid development of a national agreement regarding HPV vaccination in Japan may be primarily attributed to the advocacy of vaccine beneficiaries, supported by advocacy by celebrities and positive reporting by print and online media".

2.1.2 Social Networks and Epidemiology Studies

By growing at such quick rate, social networks are becoming valuable sources for the most varied studies and health categories also benefit from this fact. On the specific issue of tracking diseases, Yom-Tov et al. explored mass gatherings like music festivals and religious events as risky situations on transmission and spreading of communicable diseases [89]. The methodology was to use Twitter and Bing search engine to find users who mentioned one of nine major music festivals in the UK and one religious event in Mecca during 2012 and extracting postings and queries related to these events during a 30 day period after each festival. After analyzing data using a variety of methods (mainly word matching), results showed that it is indeed feasible to create public health surveillance systems for mass gatherings using Internet data.

Regarding Twitter, a few studies have taken place using this social network for tracking purposes, mainly for Influenza: Lampos et al. presented an automated tool for tracking the

prevalence of Influenza or Influenza-like illnesses in several regions of the UK; Signorini et al. focused on tracking H1N1 or swine flu in ways to allow the detection of rapidly evolution of these illnesses [44, 72]. Other interesting Twitter capability is the fact that tweets can be tagged geographically. This enables researchers to narrow the scope of their studies to a specific country and/or continent. This was what Víctor M. Prieto et al. [66] were able to do when searching for health related conditions in people (flu, depression, pregnancy and eating disorders) in Portuguese and Spanish tweets.

Internet forums are another example of preferred sources for data mining systems applied to health events: in a study by McNaughton et al., messages on internet forums are used to evaluate reactions to the introduction of reformulated OxyContin and to identify methods aimed to defeat the abuse-deterrent properties of the product [52]. Posts were collected among 7 forums between January 1, 2008 and September 30, 2013 and were evaluated before and after the introduction of reformulated OxyContin on August 9, 2010. Results showed that after the introduction of physicochemical properties to deter abuse, changes in discussion of OxyContin on forums occurred reflected by a reduction in discussion levels and endorsing content; and that analysis of Internet discussion is a valuable tool for monitoring the impact of changes in drug formulation.

2.1.3 Health Information Web Sites

Other kind of web sites focused on health information rely on information gathered from varied interactive social networks. Information related to patient experiences, drug usage and associated effects is compiled and relevant content is kept in order to build a strong repository.

Treato (treato.com) is one of the websites relying on information extracted by crawling through open health related forums, blogs and other social media sources and is more focused on drugs and medication. Information is analyzed using Natural Language Processing and then it presents aggregated patient experiences [80]. Users can search about drugs, their indications, side effects and their relation to other drugs.

A similar web site is eHealthMe.com. In this case, information is retrieved from FDA (United States Food and Drug Administration) reports and/or patient reports (it is not clear if patient reports are obtained using a proprietary social network or if those reports come from different sources) [29]. eHealthMe.com provides two different tools for information on drugs adverse side effects: drug interactions and symptom checker [1]. The first offers general drug information such as ingredients, indications and reported side effects. The second provides information on symptoms and to when they happen relatively to other patient experiences

and also about the severity and observed recovery time. More so, a question and answer forum-like service is also available.

2.2 Information Sources

Data mining projects rely on data retrieved from a multitude of sources and it is this data which will potentially contain relevant information for a certain goal or project. A data source shall suit our area of interest, have valuable information and, as such, we have to make sure we can extract exactly what we need to achieve our purposes. As stated before, the internet is a hugely relevant source for gathering information on various areas and, of course, the health area is not different. There are a few different kinds of health websites [19] that can be useful to our needs and each has a different technical approach so as how the information is provided.

2.2.1 Definition

- **Curated Sites**: sites which are sponsored by some official entity (government agency, disease-related organization or medical organization). This kind of source is very good for accessing specialized literature such as obtaining citations on biomedical literature, science journals and online books, but is not that great when our purpose is to find what the patient specific needs are regarding some condition/symptom.

  Example: PubMed.gov (maintained by the U.S. National Library of Medicine) [67]

- **Moderated Question and Answer Sites**: these sites work almost as if it was an actual doctor's appointment. Patients expose their conditions/symptoms using a text form and a medical specialist answers also in written format. For data mining, these kinds of websites are not really useful because, although all information would be very specific to what the patient needs would be, all data is kept private between patient and doctor and thus, not accessible. Example: WebMD.com [81]

- **Mailing lists and discussion boards**: this kind of interaction works in a question/answer basis, where a patient asks a question regarding a specific issue and others (patients or not) answer with advice respecting their own experience on that same issue.

  Since these kinds of sources grow fast, one of the main issues has to do with difficulties locating and accessing all relevant information.

- **Healthcare blogs**: as blogs can be written by anyone, from patients to medical specialists, the variety of subjects is huge. Some may write about health problems that someone

is currently facing, others can focus on research related to some particular subject, others can simply provide information on some disease.

The problem is, identifying health related blogs among all the others can be a difficult task. Not that there are closed access blogs, but in a way that makes it difficult to identify the specific goal of a blog automatically. For instance, we would not be interested in blogs about sports but we would probably be interested in blogs with articels on sports injuries. With thousands of active blogs, the task of choosing which to use would not be easy [2].

- **Social Networking sites**: these sites are one of the most powerful sources of information because normally they are used by a large number of people of all different social groups, ethnicities. Some of these sources are open to anyone and others are just visible and/or accessible to registered users.

  Examples: Facebook [34], Twitter [79], PatientsLikeMe [61] or Yahoo Answers [6]

2.2.2 Analysis

The process for choosing which sources to be used was defined by a few characteristics that we thought were important for this kind of work: sources should be in English language and/or its variants, should cover the largest scope of health categories we could find, should be used by a fairly number of people regardless of their geographical location or ethnicity and should favor the interaction and discussion between people.

WebMD

WebMD is a corporation known for its website which has information regarding health issues such as symptoms checklist, drugs information, blogs maintained by physicians and also has a community area where users share their experiences regarding some topic. According to Wikipedia and as of February 2014, WebMD has an average recorded 156 million unique visitors per month. This website could have been a very useful source of information for our work, but communications between patients and doctors are maintained private (as they should be) and access to the community area is restricted; only being accessible from the user point of view. These setbacks are not related only to WebMD website, but to most websites of this kind. Information is kept out of the general public access so as to ensure the privacy of the users and also to keep the information internally in ways to benefit its provider.

Facebook

Facebook is one of the most famous social networking sites: with its 874 million active users as of September 2013, this network is much preferred to gather lots of information from one single place [35]. But there is a downside on doing this: access is limited to one's connections (you can only view and/or search information from your friends - sometimes friends of friends). In some cases information is shared publicly, but this is just a small percentage of all the traffic generated by the users [71]. If we want to gather information from Facebook, we have two ways for doing so: search only public posts or build an application to which each user will grant (or not) access to certain personal information. Of course this information, if granted, would be massively useful for our research because it would be a somewhat unending source of data; on the other hand, if access isn't granted, it would limit greatly the information actually retrieved. Perhaps it would be fine to use Facebook to generate statistics on information people share with no restrictions like peoples' geographical distribution, or to study romantic relationships between friends (this has already been done by Lars Backstrom and Jon Kleinberg [11]), but regarding health issues it is fairly complicated to find something relevant. People do share a lot of information, but hardly share experiences talking about diseases, and if they do so, that information will be privately shared with a given group of friends and, as such, will not be accessible by general public [58].

Twitter

Other well known and widely used social network is Twitter. This network is rather different from Facebook: users are only able to share small messages (up to 140 characters) and they do not need to be friends to follow each other as all information is shared publicly (by default). Being used by 645 million people around the world, Twitter can be an interesting source of information: like Facebook, people tend to share a lot of information regarding the most varied issues [16]. Helping users to filter information, Twitter created the "#hashtag" concept where users can tag messages with a small reference to its purpose. Qian He et al. [68] used hashtags to find every message from Nike's fitness trackers around the globe and, after some months, were able to find interesting characteristics on peoples' sporting habits. Regarding this work, Twitter was not used because our research focus on drug names and disorders, and this kind of network is very prone to misspelled and/or abbreviated words.

PatientsLikeMe

PatientsLikeMe is a "forum like" social network focused on connecting people who share the same condition or disease. People can interact with each other, sharing experiences and track their symptoms evolution. As this network is growing (latest statistics indicate that there are more than 250 thousand active users [53]), a lot of useful information is gathered around varied conditions/diseases and, consequently, their treatment and associated drugs. This is very useful for researchers, pharmaceutical companies and regulators as a way to help future patients using information and experiences reported by past users. This network was one of our preferred to be used as an information source but we were not granted access to their database.

Yahoo Answers

Throughout the years, Yahoo has always been known to be a powerful search engine and all tasks related to access and requisition of data is somewhat easily available to the general public through APIs. One service of Yahoo's large resources (reports point to 56 million visitors in October 2013 [4]) is related to a question and answer section known as "Yahoo Answers". Yahoo Answers, created to replace "Ask Yahoo", is divided in categories (like "Arts & Humanities", "Education & Reference", "Health", "Sports" and "Travels" among others) and is available to everyone: virtually anyone with access to the internet can post a question regarding a specific issue and wait for answers to be given by other people. This way of interaction between people regardless of location, ethnicity or age, provides a very broad scope of, eventually, valuable information. One of the various categories available is health related and this makes it a perfect source for our work. After a detailed analysis, we concluded that we could retrieve all the relevant information and build a very powerful data set to fit our needs.

Reddit

Reddit is rather different to the previous social networks detailed here. Reddit has a forum like appearance where the various subjects or categories are called "subreddits"; each subreddit has an immense array of threads with statements, questions and answers. Fortunately, Reddit has a fairly large health related community comprised by 67 different subreddits (from general practice sections to more specific ones like pancreatic cancer or multiple sclerosis). Regarding users statistics, in December 2013, Reddit was accessed by 16 million unique users and had above 180 million page views [73]. Of course, these users are from all over the world which, again, can be valuable to our work, but searching data in English is more prone to retrieve data

from English-speaking countries. The main limitation of Reddit has to do with its difficult API. Even though, all information is retrievable so, Reddit was one of our choices.

2.3 Text Matching

As the main goal of this project is to be able to identify events related to health issues, one must find a solution to detect and annotate given words or sequence of words that can indicate or lead us to potential candidates for what we consider relevant. For this task, we chose text matching using dictionaries as the technique to use in order to identify words with relevant meaning. In this section, we aim to explain the methods and framework used to achieve such goal.

2.3.1 Neji

Neji [17, 57] is the main framework used for processing and annotating information. It is an open source framework optimized for biomedical concept recognition. It is focused on four crucial characteristics: modularity, scalability, speed and usability. Modularity is achieved by using an independent module for every processing task. Each module can be executed ad-hoc or integrated in a processing pipeline. This tool is able to support simultaneous applications of dozens of dictionaries and machine-learning models for concept recognition while processing large data sets: scalability. In terms of speed, the usage of concurrent processing allowing the processor cores to handle several documents at the same time is a major asset. Neji is capable of annotating up to 1200 sentences/second when using dictionary-based concept identification. Finally, as this is an open source framework, usability is a key characteristic. Developers and researchers should be able to use pre-defined pipelines, implement custom ones using provided modules and/or implement their own. Figure 2.1 illustrates the processing pipeline and modules of Neji. Core modules are sentence tagger, natural language processor, a dictionary tagger, machine learning tagger and a post-processing module. These, together with Reader and Writer modules, form the main pipeline of Neji. Neji can natively use various input and output formats, namely Pubmed XML, leXML, CoNLL and A1; if a different format is required, specific reader (or writer) modules can be created to suit. Neji achieved high performance results on named entity recognition when evaluated against CRAFT, AnEM and NCBI disease corpus (F1-measure for overlap matching are: species 95%, cell 92%, cellular components 83%, gene and proteins 76%, chemicals 65%, biological processes and molecular functions 63%, disorders 85%, and anatomical entities 82%) and on entity normalization (F1-measure for overlap name matching and correct identifier included in the returned list of

identifiers: species 88%, cell 71%, cellular components 72%, gene and proteins 64%, chemicals 53%, and biological processes and molecular functions 40%).

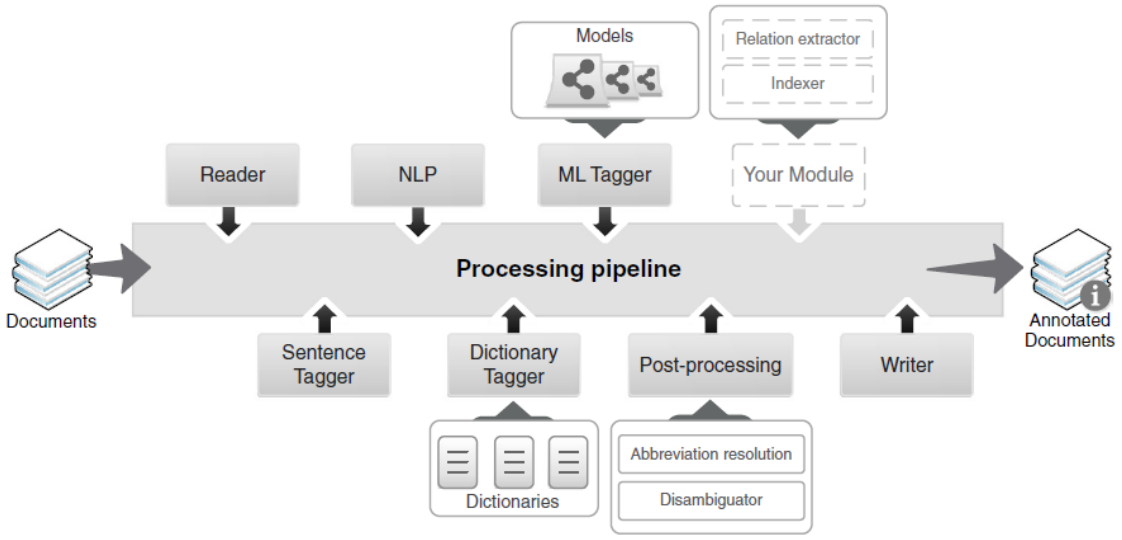

Figure 2.1: Processing pipeline and modular architecture of Neji [17, 57].

2.3.2 Dictionaries

In order to find health related events we need a comprehensive knowledge base regarding disorders, symptoms and drugs. The dictionaries used for matching disorders and symptoms were compiled from the 2012AB version of the UMLS Metathesaurus for the following semantic types [76]:

- Acquired Abnormality
- Anatomical Abnormality
- Cell or Molecular Dysfunction
- Congenital Abnormality
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Pathologic Function
- Sign or Symptom

These dictionaries were filtered in order to eliminate inconsistent names that would generate a large number of false positives: names with one or two characters, names starting with a word from a strict list of stopwords and any single word name that was included in a broader list of stopwords generated from the list of most frequent words in MEDLINE were rejected. On the other hand, relevant terms that were featured very frequently in MEDLINE such as general names of diseases (e.g. "cancer", "diabetes"), Gene Ontology terms (e.g. "expression", "transcription") and species (e.g. "human", "Saccharomyces") were removed from the aforementioned stopword list to allow their identification in texts. For identification of drug events, a dictionary was compiled using the DrugBank database [46]. After performing a few tests, we still had some false positive situations: words such as "today" were annotated because "Today" is the comercial name for a spermicide by Bliss Pharmaceuticals [49]. The solution was to clean the dictionaries once more. This time we used the British National Corpus words frequency list and selected the words which corresponded to 90% of total usage [22]. This list was then annotated by Neji and each word identified after this process was carefully analyzed. If the word was more prone to be an actual common word and not a drug or disease concept, we removed it from the corresponding dictionary; otherwise no action was taken, for instance, "aspirin" was one of the annotated words and, in this case, we did not remove it.

2.3.3 Concept Classification

After the annotation process, all concepts identified will be chemicals, disorders and symptoms. In order to classify these concepts as groups of similar purpose to allow better visualization and analysis, we compiled a series of classes for drugs and for disorders and symptoms. For disorders and symptoms we used the MEDIC-Slim classification present in the Comparative Toxicogenomics Database (CTD) [23]. MEDIC-Slim is a set of terms that classifies MEDIC diseases into high-level categories. Since we are using dictionaries compiled from UMLS, we had to cross-reference both databases for merging UMLS Id's with MEDIC Id's. For chemicals, and as we are using dictionaries compiled from DrugBank, we also compiled its list of drug taxonomy classes. Like disorders and symptoms, drugs can also be classified according their chemical function or structural properties and these kinds of characteristics might allow us to draw conclusions regarding a drug class and not only regarding a specific drug.

The full list of classes (being drug classes or disorder and symptom classes) are presented in Appendix A - Disorder and Drug Classes.

2.4 Co-occurrences

In order to analyze the collected data and to understand the connections or co-occurrences between annotations, one must use metrics to calculate the probability of an event given the presence or absence of another event. There are a few metrics used for this goal such as Odds Ratio, Relative Risk or Mutual Information among others. Although these metrics are well used in various works and studies, each one has its own positives and negatives.

2.4.1 Odds Ratio

Odds Ratio is one metric that quantifies how strongly a property A is present given the occurrence (or non occurrence) of property B. In short, the odds are the ratio of the probability that the event of interest occurs to the probability that is does not occur. According to Bland et al., Odds Ratio is widely used in medical reports because it provides an estimate for the relationship between two binary variables, it enables medical doctors to examine the effects of other variables on that relationship and it has a special and very convenient interpretation in case-control studies [14]. Gianfrancesco et al. used Odds Ratio to measure the association of antipsychotic treatments with type 2 diabetes at a population level and Gabriel et al. related the risk for adverse gastrointestinal events to the use of nonaspirin nonsteroidal anti-inflammatory drugs [39, 38]. Using a generic example to associate a chemical to a given disorder, the Odds Ratio would be computed by:

$$OR = \frac{CD \times nCnD}{CnD \times nCD} \tag{2.1}$$

being CD the number of times that chemical and disorder occur simultaneously, nCnD the number of times in which neither chemical nor disorder occurs, CnD the number of times in which a chemical occurs but the disorder does not and, finally, nCD the number of times in which a chemical does not occur but a disorder does.

2.4.2 Relative Risk

Relative Risk is similar to Odds Ratio but it is more commonly used in epidemiology and clinical trials data because its expected statistical outcome of interest has relatively low probability. Bouter et al. studied the effect of epidemic ketoacidosis, pneumonia and death in diabetes mellitus using clinical data from 1976 to 1979 [15]. Using data from the Caregiver Health Effects Study (CHES), an ancillary study of the Cardiovascular Health Study (CHS), Schulz

et al. examined the relationship between caregiving demands among older spousal caregivers and 4-year all-cause mortality and used Risk Ratio to obtain the study results [69]. Using the same generic example used to explain the Odds Ratio, Relative Risk can be calculated by:

$$RR = \frac{\frac{CD}{CD+CnD}}{\frac{nCD}{nCD+nCnD}} \tag{2.2}$$

2.4.3 Mutual Information

Mutual Information is a metric to measure the inter-variable mutual dependence. It is widely used in various areas such as search engine technology or image registration in Medical Image applications. Maes et al. applied Mutual Information as a matching criterion to find if two images are geometrical aligned or not by using statistical dependence or information redundancy between both image intensities of corresponding voxels (a value on a regular grid in three-dimensional space) [48]. Another example, this time using a variant of Mutual Information, Jeong et al. studied the possibility to assess information transmission between different cortical areas in Alzheimer's disease patients by estimating the average cross Mutual Information between EEG (Electroencephalogram) electrodes [43]. Again, using the same example as above, the Mutual Information is given by:

$$MI(C,D) = \sum_{d \in D} \sum_{c \in C} p(c,d) \log(\frac{p(c,d)}{p(c).p(d)}) \tag{2.3}$$

with p(c,d) being the joint probability distribution function of chemical and disorder and p(c) and p(d) being the marginal probability of chemical and disorder respectively.

Chapter 3

Prototype for Identification of Health Concepts in Social Networks

3.1 System Architecture

The proposed system is based on three different stages: acquiring and storing information, processing and annotating information and analysis of processed information. This section aims to explain all steps performed at each stage, and detail the challenges and solutions to overcame these. In Figure 3.1 we present a simple overview of the complete system.

As the action of acquiring information is dependent on different sources and thus different API's and different implementations, the system has to be able to process and store everything regardless of methodologies applied. To address this issue, every source has its own module to retrieve information and, regardless how every collector module works, its output is directed to an API (Data to SQL Adapter API) and thus, maintaining consistency. This way, we can always add other information sources without having to redo the necessary adapting steps or modify our repository settings. When the information reaches the Adapter API, it is then stored in the repository. This process will be further discussed in Chapter 3.2.1. The next major step in our system is the annotation process. This action is solely dependent on Neji [17] and its dictionary matching methods. After the annotation process, all concepts (and their attributes) are stored in the repository and ready to be analyzed and/or filtered.

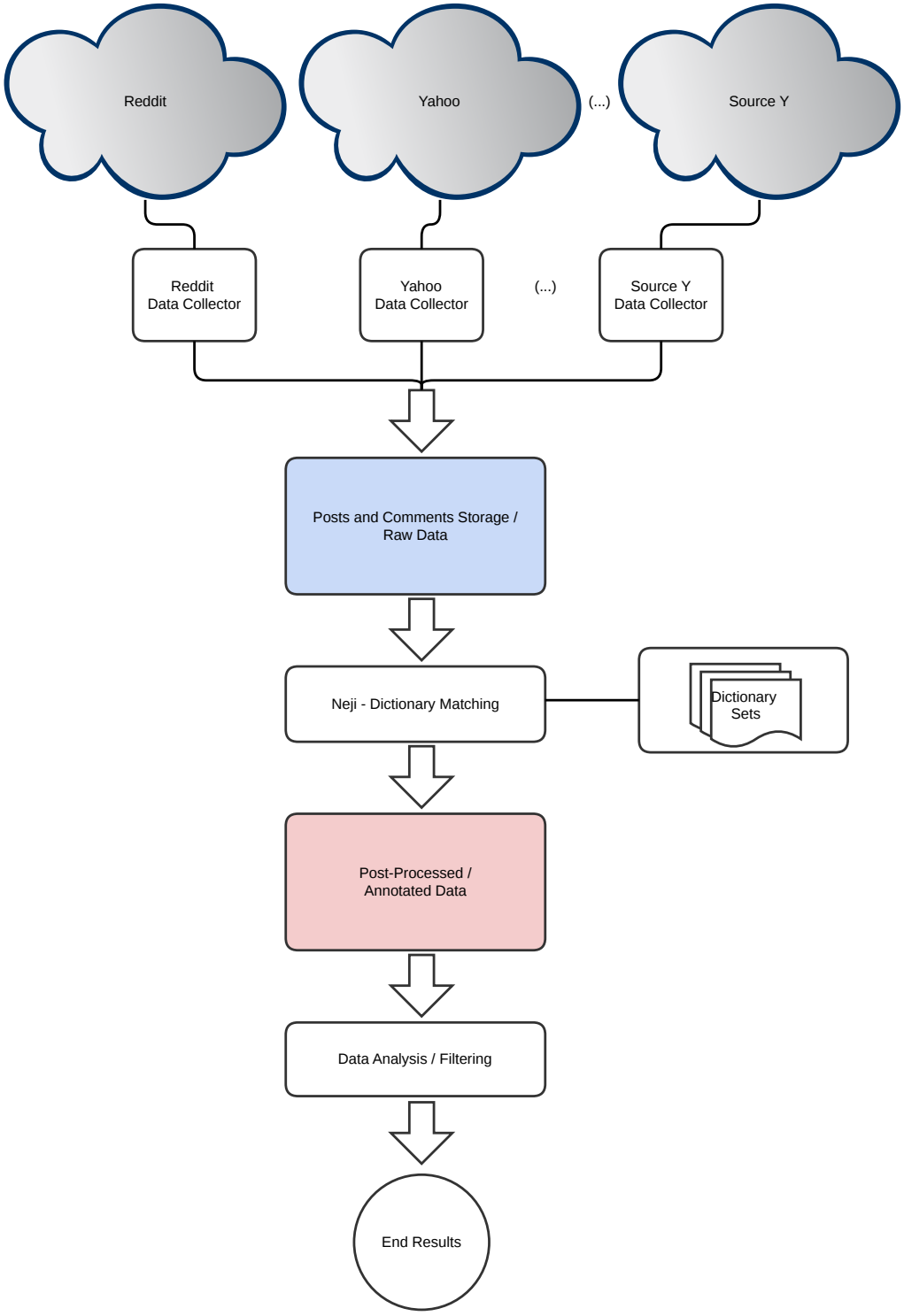

Figure 3.1: System Overview.

3.2 System Implementation

The following section describes the technical aspects and decisions made for implementing the proposed solution.

3.2.1 Database Design and Features

The main goal while designing our database was to ensure that its structure could be used regardless of the sources of information considered. If, in the future, we need or want to add a new information source, our database shall work without any problems. The database is organized in two different parts: one for storing the raw data (questions/posts and answers/replies), and other for storing the output of the annotation software (Neji). In Figure 3.2, the tables in blue are related to the first part and the tables in pink are related to the second.

The first part of the database (tables in blue) is almost self explanatory: each post and each comment has a different content; and each post and each comment originate from a given source. In order to perform a detailed analysis, in table *Post* we try to associate each entry with as many attributes as possible: each post features an *Identifier* field (this is the Id given by the source of the post) used to link the comments to the original post, a *Link* field which gives us the direct URL to that specific post, a *Timestamp* field storing the original date and time of posting; then, there are two fields acting as Foreign Keys to connect to other tables: *SourceId* to identify the post from a given source and *ContentId* to identify the entry where the actual post content (text) is stored in *Content* table. Another feature is how we store the contents of posts and comments: we use only one table (named *Content*) to do this. This decision was very important because, while annotating text, we only use this table as input without having to worry if it is from a comment or from a post (as it is, in fact, irrelevant for the annotation process). Along with the *Text* field which is used for storing raw text, there is also an *Annotated* field to be used to flag if an entry has already been processed by Neji or not (as the annotation process is asynchronous with the retrieval process, this field is valuable as explained in Section 3.2.4). Regarding the *Comment* table, every comment features the same fields as the *Post* table and two additional fields used to match posts with comments: a *PostIdentifier* which is the Id of the post this comment is associated to, and a *PostTimestamp* which is the post timestamp as a safety measure to further ensure that the comments belong to only one post.

The second part of the database (tables in pink) is very specific to our project's goals and is meant to store all information regarding concept annotation. Firstly, table *Annotation* is

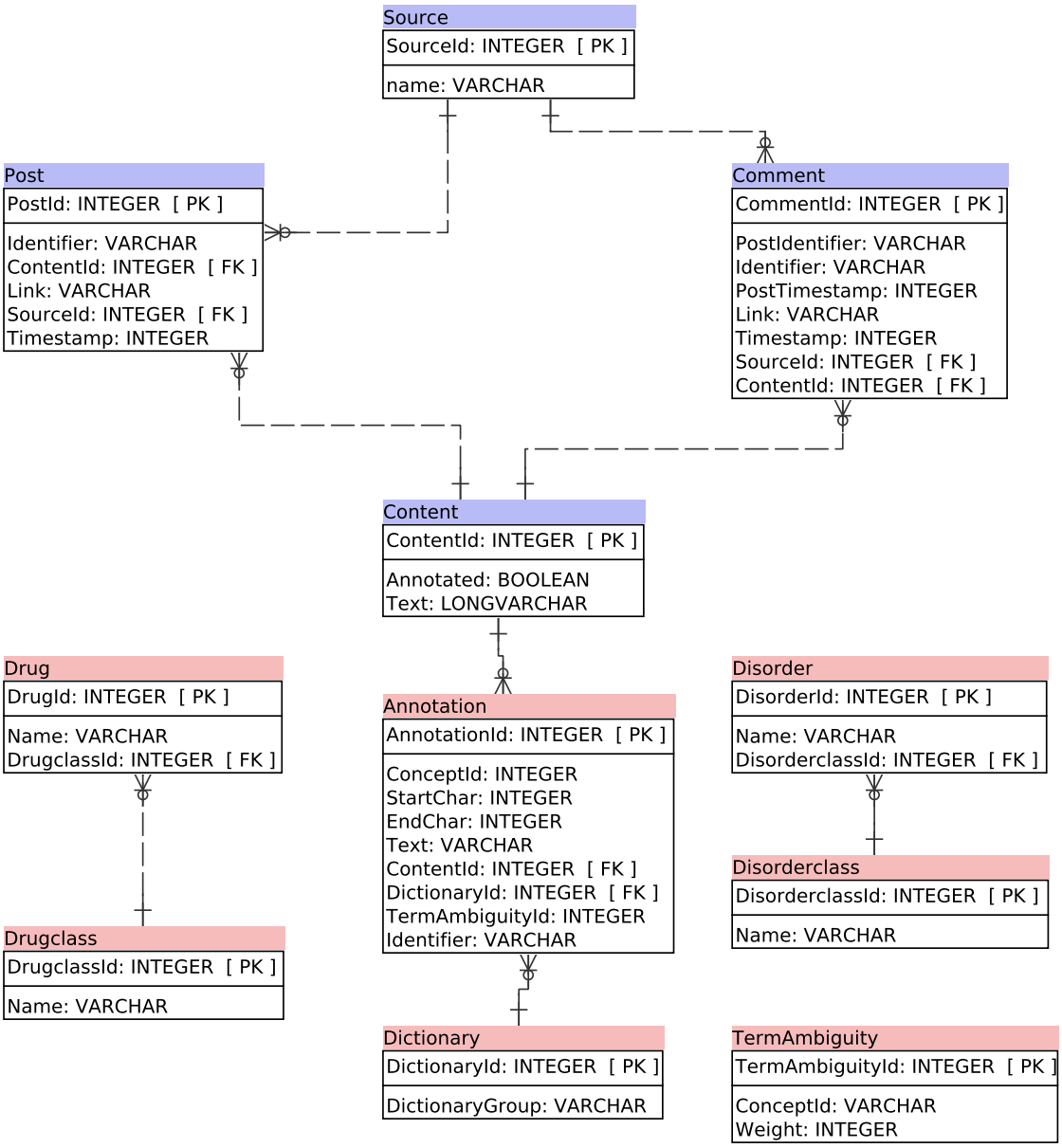

Figure 3.2: Database structure.

the core of all information regarding annotations. We need to know which content entry was annotated (using the Foreign Key *ContentId*), the mention of the recognized concept, as it appears in the original post (field *Text*), which is the position of the concept in the text (fields *StartChar* and *EndChar*) and an *Identifier* field used to store specific Neji annotation (with the format "source:identifier:type:group"). For further improvement, we added information regarding drug classes and disorder/symptom classes. As explained in Section 2.3.3, using DrugBank, we downloaded the full database and stored its contents on two tables: table

Drug which features a list of know drugs and table *DrugClass* featuring the taxonomy class of each drug, if applicable. These tables are connected by the Foreign Key *DrugClassId*. For disorder and symptoms, the procedure was identical but this time we used the Comparative Toxicogenomics Database. Drug or disorder class information has to be accessible from table *Annotation* and thus the use of *ConceptId* field. Since our project is largely based on dictionary matching, we also added a table describing which kind of dictionaries is being used and, while annotating, knowing which dictionary was responsible for each annotation. For this effect, we created the table *Dictionary* with the only purpose of enumerating the dictionaries used for each annotation. Each entry of *dictionary* table serves as Foreign Key (field *DictionaryId*) for the table *Annotation*. Finally, table *TermAmbiguity* is meant to store terms that can be found in two or more dictionary entries, that is, ambiguous terms that can be associated to more than one distinct concept in the dictionaries. The contents of this table allows us to balance each pair of concepts on our final results. If a concept has a *weight* equal to one, this means the entry *TermAmbiguityId* appears only one time in the dictionaries, and thus, has only one meaning; if a concept's weight is two (or more), this means that that concept belongs to two (or more) different dictionary entries.

3.2.2 Gathering Data

As explained in Section 2.2.2, Yahoo Answers and Reddit were the main choices to use as information sources. An analysis performed on these systems showed that both have mechanisms to supply data using JSON (JavaScript Object Notation) [42] and both systems have different sections for health related subjects and/or diseases.

Reddit - Health SubReddits

Reddit's interface, from a user's point of view, is somewhat difficult to read and, regarding its API, retrieving information is also a challenge. Although information is supplied using JSON, its query is not easy. Fortunately, there is also a subreddit dedicated to all wrappers and libraries for interaction with its interface [59]. For this task, we chose JReddit [37] which is a simple Java API for accessing Reddit's posts and comments. Similarly to Yahoo Answers, Reddit also has its Health related categories divided into 67 different subreddits, and as such, we have to perform a request for each one. Every time we request new posts, we get a reply with 25 questions and, for each one of them, we have to check if it is already stored on our database: if not, store the new ones, otherwise discard the ones already stored. Then, for each of those posts, we check for its comments and proceed the same way: if already there, do nothing; otherwise store them. As happens with a lot of systems that provide access through

APIs, we also have to respect Reddit's API request limit set to 30 requests per minute (which translates to 1800 requests per hour) [8]. From our tests, we perform 67 requests for each category (which returns 25 posts) and perform a request for each of those posts performing a total of 1675 requests per hour.

A more detailed procedure to fetch posts and comments is as follows: using JReddit API, the first step for retrieving posts and/or comments from Reddit is to get an instance of the Reddit API. Once we do this, we can use that same instance for getting posts from a specific subreddit. We are interested in 67 different health related subreddits so we must create an array with all subreddit specific keywords (such as "health", "adhd" or "cancer") to be used as input for each different retrieval procedure. As we start requesting posts from each subreddit, we can choose the way data is retrieved: new posts first, top posts first, controversial posts first, etc. For our purpose, we are only interested in new posts. After requesting new posts from a given subreddit, the data returned is passed into a Listener which has two functions: onSuccess and onFailure; both names are pretty self-explanatory. In case of success, we will now have a list of 25 posts from which we retrieve the relevant information from the post such as identifier, contents, link, timestamp and we store this information in an ArrayList [10]. While we are navigating through the posts, we are checking if there are any comments associated. If so, we retrieve their content, id, timestamp, link and store everything in a separate ArrayList. The usage of ArrayLists to store posts and comments has to do with an architectural decision which will be explained a bit further.

Yahoo Answers - Health category

Retrieving data from Yahoo Answers can be performed like a SQL query (using the appropriately named YQL - Yahoo Query Language [45]) via HTTP Request to the Yahoo Query API. For example, the following YQL query would retrieve the newest 50 resolved questions posted on the Health category of Yahoo Answers:

```
select * from answers.getbycategory(0,49) where category_id= 396545018 and type="resolved"
```

Queries performed using YQL return data in JSON format and, to be able to retrieve the objects we aimed for, we had to build an interpreter from scratch. Gson [40] was the chosen library to convert JSON data into java objects for us to use with ease. But after trying a few queries, some would work well and create perfect objects and others would not work at all. The problem was when a given question had only one answer: in case of multiple answers, the JSON reply would feature an array with the answers but in case of a single answer, JSON reply would have a single object and not an array with only one object as we thought. To overcome

this issue, we built two different Gson object classes and a simple "try ... catch" mechanism, one to deal with single answers and other to deal with multiple answers. The next step is to perform a series of YQL queries to retrieve 50 lists of 50 questions for each of three different question categories: open questions, resolved questions and undecided questions. Queries are performed on category id 396545018, which represents the id of the Health category. For each of the questions retrieved, we store it in an ArrayList. Of course, replies or comments to each answer are valuable to us, so whenever we store a new question we also perform a request for its answers. Then again, the retrieved answers are stored in a different ArrayList. It is important to mention that, to be able to freely query Yahoo's database, we are limited to 2000 queries per hour [9]. This limit could be raised if we created an account, but in case we lost those credentials, our system would end up useless so we chose not to go that way. In any case, retrieving questions hourly, we perform around 250 queries which are well under the recommended limit.

3.2.3 Storing Retrieved Data

At this point we had two choices for storing retrieved data on our repository. The first one relied on each data collector to be freely allowed to access the repository and perform different tasks with the stored data and, obviously, to store the contents of each collector run. This solution, although a bit quicker and easier to understand, could lead to potential data losses if some operation was not done carefully. This would also means that in order to build a different module for a different data source collector, the person would have to have a certain knowledge as to how the repository and its database is built. On the other hand, the second solution would be to create a step between collecting and storing data; that is, data still is retrieved by each collector but instead of storing it directly, it is sent in the form of a list to the intermediary module to be, then, stored in the repository. With this solution every new module would only have to be focused on retrieving data and, above all, if some changes were to be made in the repository, the only modifications one need to do were to be performed on the intermediary module leaving the collector modules untouched.

We decided to use the second option hence the usage of ArrayLists on the collector modules detailed in Section 3.2.2. For this solution to work, we must provide certain directives to be followed when building data collector modules:

- ArrayLists must be composed by objects Post or Comment (as specified in Figure 3.3)
- In the end of the retrieval process, ArrayLists should be passed to the DataToSqlAdapter module (Figure 3.4)

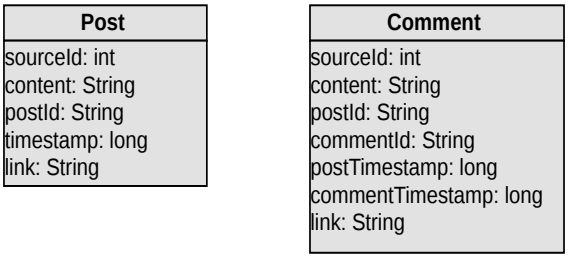

Figure 3.3: Post and Comment objects structure.

```
public class MyDataCollector {

   public static void main(String[] args){

      ArrayList arrayPosts = new ArrayList();
      ArrayList arrayComments = new ArrayList();

     // Data Collector API specific code

     // Inserting a new Post Example
      arrayPosts.add(new Post(0, "content", "125jjzz",
               123456789, "http://link.to/post"));

     // Inserting a new Comment Example
      arrayComments.add(new Comment(0,"content","125jjz",
              "125jjz1",123456789,123456790,"http://link.to/comment"));

     // Calling Adapter
      DataToSqlAdapter(arrayPosts, arrayComments);

   }
}
```

Figure 3.4: Simple example code for a data collector

On receiving lists of posts and comments, the DataToSqlAdapter module will then perform a series of steps in order to store the received data. The module starts a new SQL connection with the database and then, while processing each comment and each post, verifies if it is already present on the database or not: if already present, the module will discard; otherwise, the comment or post is stored in its respective location. Upon ending processing all comments and posts lists, the module will terminate the SQL connection.

3.2.4 Concept Annotation

Since our system is now able to acquire and store data retrieved from Yahoo Answers and Reddit, the next step is the annotation phase. In this section we explain which steps were taken to annotate the aforementioned data.

Neji - Modifications to allow SQL interaction

Neji is an essential tool in our system as it is used to analyze and annotate raw data retrieved from Yahoo and Reddit. When using Neji's command line interface, the pipeline was built to use a batch of text files and work on those files; and, after performing its analysis, the output is also stored in text files with a given format. This had to be overcome, otherwise each time we had to analyze a batch of posts or comments, we would have to write all data to text files, feed them to Neji; after Neji analyzed everything, we would have to read all output files and store data in the database. As this was not a practical solution, we decided to modify Neji to allow SQL interaction all by itself. We had two areas we needed to work on: how to read data from the database and how to store it back on the database.

Neji - SQL Read

As explained above, for our solution choice, Neji's pipeline supports large batches of text files as input in which the name and contents (corpus) of the file are essential: the name for identification purposes; and its contents whose raw data is to be used as corpus and, later, analyzed. With that in mind, we had to find a solution to make sure these two variables were always present when using a database as input source and that this same solution is as flexible as can be (in case of future updates to the system). This challenge was solved by creating a *SQLBatchExecutor* (highly inspired by the *FileBatchExecutor* already developed which is where the processing pipeline is started) and by creating a configuration file (appropriately named *sqlproperties*) in which the user specifies the hostname, login, password and database from which data is to be retrieved. There is also another information that needs to be specified in that same file: a SQL Query whose format has to obey a structure where the first column values are used as identifiers and the second column values are used as corpus input. The *SQLBatchExecutor* performs exactly the same steps as the *FileBatchExecutor*: creates the pipeline processor, initializes context and starts thread pool, processes file entries (in the case of *SQLBatchExecutor*, processes SQL entries) and is also responsible for ending all operations in the end. The main difference between these two executors is that one relies on a batch of files as input source and the other uses a SQL query (present on the configuration file named above) to retrieve needed data as input. Obviously, these two mechanisms are to be used at different times, and this is guaranteed by using different switches when starting a new command-line job.

sqlproperties example

```
hostname    = 127.0.0.1
database    = <database>
username    = <user>
password    = <pass>
inputQuery  = SELECT contentid,text FROM content WHERE NOT annotated ORDER BY contentid ASC;
```

Other problem that needed solving was the fact that Neji had no support for SQL, so we had to develop some kind of manager to perform actions on the database without interfering with what was already done in Neji itself. To address this issue, a module was developed to allow for SQL management (which we named *SQLManager*) so as to perform actions like connect to and disconnect from the database, perform selects, inserts and other actions essential to the project. This module was developed as a singleton [77] to allow a large number of SQL operations using only one connection to the database (otherwise we would need hundreds or thousands of connections according to the volume of data to be analyzed and the database would just refuse any more connections than a given number). Being a singleton, the instance of *SQLManager* is created when the first connection is needed (namely when entries to process are retrieved from the database using the query specified on the configuration file); afterwards, all other operations use the same instance of *SQLManager* to perform its actions. In the end, after the analysis is finished, the connection is terminated and the instance is destroyed. Now, with the input data problem solved, we needed to find a solution for how to deal with the output.

Neji - SQL write

Being highly modular, Neji is a very good tool to support a lot of enhancements and one of its richest features is the ability to support third-party modules to achieve whichever goals we need to meet. Neji has already built-in modules that enables us to store analyzed data in various formats such as XML, A1 [5], JSON or even a Neji format so, for our specific goal, we could use one of these formats and simply modify it. We chose to use the Neji A1 [17] format because it has all the information we need like annotation type, start and end offset of given annotation; and, above all, for debugging purposes when developing our database driven solution, it is very easy to read and understand. Maintaining the trend when developing the modules responsible for data output, we named our module *SQLWriter*. Our module is derived directly from the *A1Writer* module already present in Neji and the main differences are the data which we want to store and the way we store information (SQL Insert vs file write). We decided to store the following attributes: dictionary group, start char and end char of given annotation, the annotation text, corpus id and the Neji specific dictionary identifier. This

identifier is provided in the following format "source:identifier:type:group", and in case there are more than one identifier per annotation, a pipe ("|") is used to concatenate all identifiers. "source" is the dictionary source, "identifier" is the identifier inside the dictionary, "type" is the semantic type (if available) and "group" is the semantic group. So, for each time that *A1Writer* would perform a file write operation, our module performs a SQL Insert operation. The way of working is exactly the same but each has its different output system. This module has the problem of being a bit specific to our project because we want Neji to perform rather specific tasks regarding our database design. Although we managed to avoid this problem on the SQL Reader part, in the case of the SQL Writer, the solution was a bit more rigid.

Since our project has some needs that are very specific to our case scenario, there are two operations which were built to deal with very specific problems we faced while developing the system. One has to do with the text retrieved from Yahoo Answers and Reddit not being normalized which would translate in problems for our post-annotation analysis. Although Neji does not have any problems with this because while matching concepts it is not case sensitive, while writing the output it keeps the initial format and that can lead to problems while performing statistical analysis. For example, if an annotation is written in uppercase characters and that same annotation is written elsewhere in lowercase characters, our repository will interpret these as two different annotations (because it is case sensitive) when, in fact, they are the same. So, in order to overcome this problem, before any data is passed to Neji, all text is changed to lowercase characters. This is achieved with a specific SQL Update Query using a postgreSQL native function.

The second problem is to flag all annotated text so as not to be annotated again in a future Neji run. Text content is always being stored and our database is growing hourly (every time that our Reddit and Yahoo Answers data collectors are activated), so whenever the annotation process begins, the system needs to know which entries were already processed and ignore them. The way we chose to deal with this situation was to add the boolean field *annotated* to the Content table which is TRUE when that entry was already annotated or FALSE otherwise. For this solution to work, we need to update the status of the processed entries to TRUE.

3.2.5 Analysis and Filtering

With data correctly annotated and with concepts stored in the repository, the next logical step is to analyze the contents achieved. Being the main goal of this project to find associations and co-occurrences between concepts, we had to find a mechanism to co-relate concepts annotated on posts with the ones annotated on their comments. The solution was to order all concepts

using the id of the post it belongs to (remember that also the comments are linked to the original post by the post id). With this solution, and by processing post by post, we achieved a list of co-occurrences between concepts. Using Python to process lists in an efficient way, we compute the Odds Ratio between each pair of concepts (being drug/drug, drug/symptom, drug/disorder, disorder/symptom or disorder/disorder) to obtain five lists with our almost final results. The last step is to filter the somewhat obvious results. If our goal is to find different and/or unexpected relations between drugs and disorders, it makes sense to remove results that are somewhat obvious or well known to the medical community and, hence, not interesting for our purposes. To filter the results, we extracted a chemical/disease associations list from the Comparative Toxicogenomics Database (source also used for disorder classification as explained in Section 2.3.3) to serve as filter [27]. The final step was to remove all associations present in this list from the drug/chemical and drug/symptom co-occurrence lists we had obtained previously, leaving us with the final results.

The way we engineered this step was to dump the information related to the annotated concepts to a CSV (Comma-separated values) file and use Python routines to perform the steps explained above. In the case of the filtering step, we used another CSV file to contain the known chemical/disease associations so as to enable us to add other associations that we see fit or that are missing from the original list from where we compiled this one. The sole requirement to add associations to this list is that the disorders or symptoms must be specified using UMLS identifiers and drugs must be so using DrugBank identifiers.

3.2.6 Main Control Module

Autonomy being one of the main characteristics of this project, there has to be a control module always running to ensure each task is performed when should. As such, the control module is responsible for triggering the data collection and concept annotation processes. As explained in Section 3.2.2, each data collector must not exceed a certain number of API requests during a given period of time and the only way to do this, is to carefully schedule each data collection run. To be able to perform these tasks, the main control module relies on using cron4j which is a powerful scheduler for Java quite similar to the UNIX cron daemon that is capable of triggering java methods as well as system processes and/or sh/bat scripts if needed [20, 21]. Similarly to the annotation process, the main control module also makes use of a configuration file that specifies which task must be run at which time.

Main Control Module properties example

```
# Run Reddit collector each hour at minute 10
Reddit = 10 * * * * *
# Run Yahoo collector each hour at minute 20
Yahoo = 20 * * * * *
# Run annotation process each week on sundays at 22h05m
Neji = 5 22 * * 0
```

3.2.7 Technologies

Java

Since our main goal was to develop an entire solution for fetching, storing and processing all information automatically regardless of platform, the chosen language was Java. With Java, we have the ability to change platforms, if needed, without having the trouble of porting the entire code; we have an infinite resource of documentation because many developers around the world use Java and already have some open source solutions that might be useful to us; also, Java prevents incorrect behavior of our code to affect the rest of the computing environment by running on a virtual machine. Other reason to chose Java, had to do with us wanting to re-use a framework previously developed by our group that also uses this technology: Neji [17, 57].

PostgreSQL

One of the main issues with data mining systems is related to the capability for storing large amounts of data. It is not unusual to have repositories with millions of entries and the system has to be fast enough to ensure quick operations on or with that large amount of data. It is absolutely crucial to have a solid database engine behind our repository to ensure that, when analyzing data, all is consistent and correct so as to infer situations, behaviors and other critical variables which could be of our interest. With consistency being our main target, choosing PostgreSQL [65] was somewhat easy: PostgreSQL community is known for always placing data integrity on top of their priorities. PostgreSQL is well known for its Multiversion Concurrency Control (MVCC) that ensures ACID principles (Atomicity, Consistency, Isolation and Durability) in an efficient way: each transaction is given a "snapshot" of the database allowing changes to be made without being visible to other transactions until the changes are committed. This solution also eliminates the need for most read locks usage [64]. PostgreSQL also supports international character sets, multibyte character encodings, Unicode and it is

locale-aware for sorting, case-sensitivity and formatting which is important when dealing with data retrieved from sources available to all countries in the world. Regarding maximum limits, PostgreSQL is more than capable to ensure our needs as shown in Table 3.1.

Limit	Value
Maximum Database Size	Unlimited
Maximum Table Size	32 TB
Maximum Row Size	1.6 TB
Maximum Field Size	1 GB
Maximum Rows per Table	Unlimited
Maximum Columns per Table	250 - 1600 (depending on column types)
Maximum Indexes per Table	Unlimited

Table 3.1: General PostgreSQL limits. [65]

Chapter 4

Results and Discussion

In this chapter we will analyze gathered data and show which relations were obtained and what conclusions we got from them. For the following statistical analysis we used all information collected since October 2013 to June 2014.

4.1 Data Characterization

The two sources used, Reddit and Yahoo Answers, proved to be really active sources of information as shown in Figure 4.1. Since this project began in September 2013, we were able to start gathering data from October onwards. October and November were used to fine tune our data collectors and, as such, the information gathered was very small but, nonetheless, valuable. In those first two months we gathered about 850 posts and 2100 replies which resulted in, more or less, 5400 annotated concepts. In December we had already gathered a rather large amount of information, but from January 2014 onwards our system started to achieve its full capacity, peaking at 73800 posts, 100100 replies and 239425 annotated concepts in April alone. From October 2013 to June 2014, we collected information from 311000 posts, 528000 replies resulting in 1168000 concepts. During the 9 month period, this corresponds to 1150 posts, 1950 replies and 4320 concepts per day.

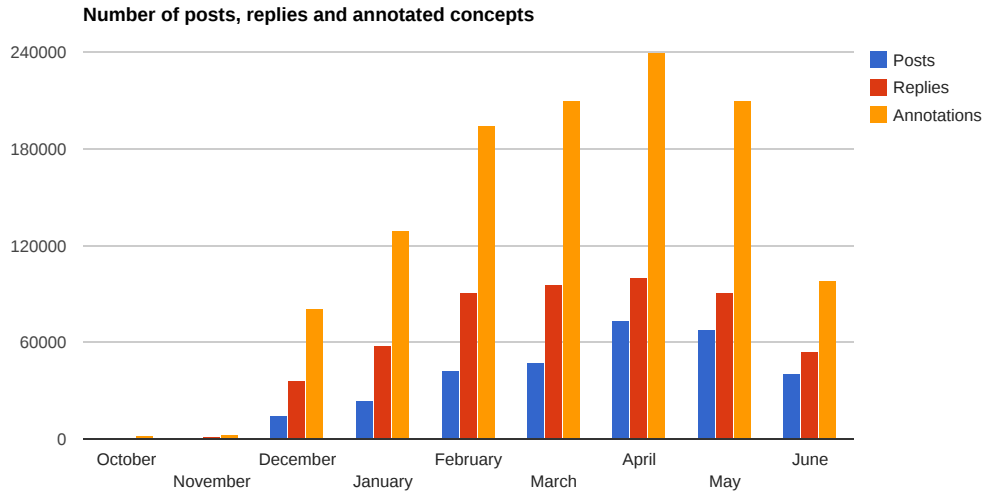

Figure 4.1: Number of posts, replies and annotated concepts.

4.2 Top Annotated Classes

A more detailed analysis performed over what concepts were annotated, allows us to pinpoint which health categories were more active during this 9 month study. The contents of Table 4.1 are the top 10 classes as extracted from the repository.

Looking at the drug column, steroids and its derivatives are the first occurrence in the list which means that these kind of drugs the ones with most references in our samples. In any case, steroids are the same or similar to certain hormones present in the human body which are responsible for fighting stress and promoting growth. It is not unusual to use steroids in various formats such as pills, gels or injections for varied medical treatments such as reducing cholesterol levels and for improving sports performance [74]. Next in the list are the amino acids. Amino acids play a crucial part in the human body: they are present in our cells, muscles and tissues and carry important functions such as giving cells its structure, for example. Amino acids are also crucial in the transport and storage of nutrients and also influence organs, glands, tendons and arteries related functions. Healing wounds and repairing tissues in bones, muscles and skin are among the most important roles of amino acids citeaminoacids. The next class are the alkaloids and its derivatives. Alkaloids are chemicals of plant origin composed of carbon, hydrogen, nitrogen and oxygen and have pronounced effects on humans' and animals' nervous system. Some commonly known alkaloids are caffeine, nicotine, cocaine or morphine [75]. In fourth place comes the xanthines class. This kind of drug is used as medication to improve

	Drug		Disorder	
1	Steroids and steroid derivatives	1871	Signs and symptoms	187205
2	Amino acids	1619	Mental disorder	151881
3	Alkaloids and alkaloid derivatives	1423	Nervous system disease	120969
4	Xanthines	1202	Respiratory tract disease	55328
5	Phenylacetates	1038	Skin disease	50656
6	Phenols and derivatives	719	Pathology (process)	50166
7	Inorganic ions and gases	612	Cancer	37857
8	Phenylpropylamines	440	Viral disease	36943
9	Benzofurans	420	Bacterial infection or mycosis	35708
10	Benzodiazepines	397	Musculoskeletal disease	28485

Table 4.1: Top Annotated Drug and Disorder Classes.

respiratory function by opening air passages in the lungs. Some of its indication is for the treatment of asthma, bronchitis and emphysema [24]. The last drug class we will mention are the phenylacetates. This class of drugs are mostly associated to aspirin and its derivatives and it is not unusual to appear on the top of our list [3].

Now looking at the disorder column, these classes are almost self-explanatory so we will just point some curious aspects of this side of the list compared with what we explained above. It is no surprise to see signs and symptoms being at the top of the list, because the main focus of every person who asks for some kind of medical help, will be to list signs and symptoms experienced. There is an obvious relation between drug and disorder classes because if a certain disorder is referenced, drugs associated with its treatment will also be referenced. For instance, if xanthines are used to improve respiratory function, they will be used for respiratory tract diseases.

The complete lists of drug and disorder classes are present in Appendix A - Disorder and Drug Classes.

4.3 Verification and Validation of Results

In the next subsections, we present relations between drug/disorders and drug/symptoms that were found through our analysis and that are previously known to exist (as explained in Section 3.2.5). This is important to validate our results and, as such, allowing us to assume the remaining associations are correct even if unexpected. For this task we rely on DrugBank

and the drug indication for each of the drugs listed below.

More complete tables can also be found in Appendix B - Lists of known concept co-occurrences.

4.3.1 Known Co-Occurrences between Drugs and Disorders

	Drug	Disorder
1	cyanocob(III)alamin	Anemia
2	metformin	Polycystic Ovary Syndrome
3	metformin	Insulin Resistance
4	cortancyl	Crohn Disease
5	L-thyroxine	Hypothyroidism
6	clomiphene	Polycystic Ovary Syndrome
7	testosterone	Erectile dysfunction
8*	thiamine(1+)	Autistic Disorder
9*	calciol	Autistic Disorder
10	tramadol	Fibromyalgia
11	progesterone	Infertility
12	metformin	Diabetes Mellitus, Non-Insulin-Dependent
13	sertraline	Paranoia
14	clomiphene	Spontaneous abortion
15	diphenhydramine	Urticaria

Table 4.2: Known co-occurrences between drugs and disorders.
* - denotes indirect relation

Since the majority of results are direct associations between drug and disorders, we will not analyze them line by line but will, instead, focus on lines number 8 and 9 as these are not immediate relations. According to DrugBank, Thiamine(1+) (commonly known as vitamin B1) is used for treatment of thiamine and niacin deficiency states, Korsakov's alcoholic psychosis, Wernicke-Korsakov syndrome, delirium and peripheral neutritis. On the other hand, children with autism normally suffer from nutritional deficiencies and their diet is normally complemented with vitamin supplements [36]. The previous statement also validates the association present in line 9 as calciol is used for treating vitamin D deficiency.

	Drug	Symptom
1	cortancyl	Arthralgia
2*	venlafaxine	Xerostomia
3*	methylphenidate	Xerostomia
4	famotidine	Dyspepsia
5	melatonin	Asthenia
6	progesterone	Hot flushes
7	cyanocob(III)alamin	Dizziness
8	potassium	Spasm
9	famotidine	Heartburn
10*	bupropion	Xerostomia
11*	(S)-duloxetine hydrochloride	Xerostomia
12	morphine	Abdominal Pain
13	melatonin	Fatigue
14	aripiprazole	Agitation
15	quetiapine	Agitation

Table 4.3: Known co-occurrences between drugs and symptoms.
* - denotes side effect relation

4.3.2 Known Co-Occurrences between Drugs and Symptoms

All the results in Table 4.3 can be verified using DrugBank except 2, 3, 10 and 11. Common to the lines mentioned is xerostomia, a condition characterized by dryness in the mouth. Xerostomia symptoms are commonly associated to side effects of different types of medication such as anti-depressants or antihistamines. Following the analysis of venlafaxine, methylphenidate, bupropion and (S)-duloxetine hydrochloride, we found these are, in fact, anti-depressants. So in the cases of the lines listed above, the symptoms occur from side effects to certain types of medication.

4.4 Statistical Analysis

All tables in this section result from the computation of the Odds Ratio between each pair of concepts (Tables 4.6, 4.7 and 4.8) and after filtration of the known drug and disorder/symptom associations (Tables 4.4 and 4.5). As with the previous section, the full result tables are presented in Appendix C - Lists of concept co-occurrences.

4.4.1 Co-Occurrences between Drugs and Disorders

	Drug	Disorder
1	caffeine	caffeine stimulant related disorder
2	glipizide	Prostate carcinoma
3	glipizide	Malignant neoplasm of prostate
4	miconazole	Yeast infection
5	glipizide	Memory impairment
6	sertraline	Self hatred
7	cyanocob(III)alamin	Chronic Fatigue Syndrome
8	infliximab	Crohn Disease
9	sertraline	Loss of interest
10	sildenafil	Premature Ejaculation
11	infliximab	Cancer Remission
12	adalimumab	Crohn Disease
13*	glipizide	Psoriasis
14*	glipizide	insanity
15	clomiphene	Infertility

Table 4.4: Top co-occurrences between drugs and disorders.
* - denotes inconclusive relation

By analyzing Table 4.4, we may find some unexpected associations between drugs and disorders. Again, DrugBank will be the source of pharmacology indication and toxicity. The first line associates caffeine with caffeine stimulant related disorder. This relation is obviously triggered by excess of caffeine consumption. The following lines will be grouped together as glipizide is the common drug present in the associations: 2, 3, 5, 13 and 14. Glipizide, is used as a diet adjunct to control hyperglycemia in patients with diabetes type II and, as such, not evidently related to any of the disorders listed. However, according to eHealthMe.com, glipizide has a rather large list of reported side effects with Prostate carcinoma, malignant neoplasm of prostate and memory impairment being listed as known side effects [32]. Although

patient reports indicate these relations, its validity may prove to be incorrect until recognized by an health entity or agency. For psoriasis and insanity, there are also a few reported cases but in smaller proportion. These are the kind of relations we want and are the ones which need to be scientifically tested so as to access its accuracy. And, to our knowledge, these results worth further exploration. Moving on to lines 6 and 9 which share the same drug: sertraline. Sertraline is a drug used in the management of major depressive disorder, so being associated with self hatred and loss of interest which in turn are common symptom for personality disorders. The next association is between cyanocob(III)alamin (vitamin B12) and Chronic Fatigue Syndrome. Chronic Fatigue Syndrome is characterized by tiredness not due to exertion, not significantly relieved by rest, and is not caused by other medical conditions. Although this disease has no cure, symptoms can be managed or treated using anti-depressants and anti-anxiety drugs. Other forms of avoiding symptoms rely on healthy diets and vitamin supplements to improve overall energy [60]. Vitamin B12 is included in the list of vitamins used, but its efficiency is not clear to help patients suffering from Chronic Fatigue Syndrome. The next two co-occurrences share the same drug, infliximab, but for rather different disorders, Crohn's disease and cancer remission (lines 8 and 11). Being the described disorders rather different one could suspect their relation but, in fact, infliximab is used for treating signs and symptoms of Crohn's disease and ulcerative colitis. For cancer remission, according to Bickston [13], it is reasonable to speculate that infliximab, while used for treatment of ulcerative colitis, may decrease the risk of colon cancer. Line 10 is an interesting case of placebo effect: sildenafil (commonly known as Viagra) is indicated for erectile dysfunction disorders but not premature ejaculation. Although this relation can be wrongly assumed, McMahon et al. [51] stated that sidenafil increases confidence, overall sexual satisfaction and increases the perception of ejaculatory control, thus, creating placebo effect. To end the analysis of this table, lines 12 and 15 are examples of direct relation between drug and disorder: adalimumab is used for treatment of a variety of disorders including Crohn's disease; and clomiphene is used to induce ovulation in female patients suffering from infertility related to anovulation.

4.4.2 Co-Occurrences between Drugs and Symptoms

	Drug	Symptom
1	cyanocob(III)alamin	Trembling
2	sildenafil	Memory Loss
3	glipizide	Memory Loss
4	infliximab	Flare
5	cyanocob(III)alamin	Weakness
6	glipizide	Back Pain
7*	melatonin	Trembling
8	adalimumab	Flare
9	cortancyl	Flare
10	cyanocob(III)alamin	Lassitude
11	glipizide	Sleeplessness
12	glipizide	Stomach ache
13	cyanocob(III)alamin	Asthenia
14	testosterone	Blurred vision
15	glipizide	Ache

Table 4.5: Top co-occurrences between drugs and symptoms.
* - denotes overdose relation

Similarly to table 4.4, glipizide is also featured on this list in lines 3, 6, 11, 12 and 15. As stated before, glipizide is prone to have different side effects and memory loss was one of the identified earlier. Back pain, sleeplessness and ache (or pain) are also among the side effects mostly identified by patients. Regarding stomach ache there are no reports stating this symptom although a few patients report nausea, vomiting and gastrooesophageal reflux, so this particular association can be plausible. On the second row, we find sildenafil (viagra) and memory loss. As explained earlier in this document, sildenafil is used for erectile dysfunction disorders but memory loss is also documented to be related to viagra by Savitz et al. [70]. The authors state that "exposure to sidenafil in migraineurs may trigger vasomotor instability, which in turn could lead to TGA (transient global amnesia) by presumably impairing hippocampal function" and, according to WebMd, all three erectile dysfunction drugs list amnesia in their labels since 2008 [82]. For line 4, infliximab (or remicade), a drug for treatment of Crohn's disease is associated with flare. Flare is defined by "a sudden intensification of a disease" or being "an area of redness on the skin surrounding the primary site of infection or irritation" [25]. If the first definition is rather subjective, the second can be related to a side effect reported by patients when using infliximab [31]. The following association is line 7 and

relates melatonin with trembling. Melatonin is an hormone which helps regulate sleep-wake cycles and is used as an oral supplement for jet lag, insomnia and other sleep related disorders. Although being an hormone naturally present in the human body, if overdosed can result in drowsiness, headache, mild anxiety, tremors, among other symptoms. Lines 8 and 9 relate different drugs for the same symptom. Adalimumab and cortancyl are drugs related to flare in the form of skin rashes. Similarly to line 4, adalimumab is also used to treat Crohn's disease and is also stated to cause skin rashes as a side effect [30]. However, cortancyl is used for the treatment of drug-induced allergic reactions, systemic dermatomyositis and other skin related disorders so, in this case, its relation with flare is direct and not a side effect symptom. Finally, line 15 associates testosterone with blurred vision. Testosterone is an hormone which stimulates sexual behaviors and sexual drive, promote secondary sexual characteristics such as increased muscle, bone mass, and the growth of body hair, and is essential for health and well being [88]. Testosterone is also used for managing hypogonadism (a disorder caused by low testosterone levels) and andropause. Patients who do not naturally produce enough testosterone are normally treated with testosterone replacement therapy and, in some cases, some side effects occur: rapid or slow heartbeats, nervousness, blurry vision, pain in the bladder, pelvis or stomach and headaches just to name a few [47].

4.4.3 Co-Occurrences between Drugs

	Drug	Drug
1	thiamine(1+)	calciol
2	infliximab	adalimumab
3	cortancyl	infliximab
4	amphetamine	dextroamphetamine
5	lamotrigine	lithium tetrahydroaluminate
6	adalimumab	cortancyl
7	lamotrigine	aripiprazole
8	lisdexamfetamine	dextroamphetamine
9	aripiprazole	lithium tetrahydroaluminate
10	lisdexamfetamine	amphetamine
11	oxycodone	morphine
12*	clomiphene	progesterone
13	quetiapine	aripiprazole
14*	glipizide	testosterone
15	quetiapine	lamotrigine

Table 4.6: Top co-occurrences between drugs.
* - denotes indirect relation

For the following analysis and discussion, we will use DrugBank as our main source of information regarding drugs and its pharmacology indications. Starting with thiamine(1+) and calciol, both are used for treating vitamin deficiencies: thiamine (vitamin B1) and niacin (vitamin B3) for the former, and vitamin D for the latter. On a naive approach, we will validate this association as mildly expected. The 2nd, 3rd and 6th rows feature drugs used in the treatment for Crohn's disease and, as such, these associations are also correct and also expected. The next associations to be discussed are present in lines 4, 8 and 10. Associations between amphetamine and dextroamphetamine which we will also relate to lisdexamfetamine. These relations are curious because dextroamphetamine and lisdexamfetamine are indeed amphetamines and amphetamines, in general, are used to treat attention deficit hyperactivity disorder. Again, these associations are correct. Lines 5, 7, 9, 13 and 15 are also interconnected. Lamotrigiene, lithim tetrahydroaluminate, aripiprazole and quetiapine are drugs used for treatment of mental disorder illnesses such as schizophrenia, depression and bipolar disorder. These drugs are also associated correctly. Moving to line 11, oxycodone - a semisynthetic opiate used for treating diarrhoea, pulmonary oedema and for the relief of moderate to moderately severe pain - and morphine - a well known narcotic pain management agent used for relief

and treatment of severe pain - is the next association. In this case, it is somewhat obvious their relation as both being used for pain management. The next association is a bit more difficult to access its accuracy. On line 12 we have clomiphene, a drug for inducing ovulation, and progesterone, an hormone naturally secreted to control menstrual cycle, pregnancy, and embryogenesis in humans. When lacking progesterone, pregnancies are not possible and in these cases, progesterone levels can be regulated by using supplementation thus allowing the aforementioned pregnancies on previously infertile women. The relation between these two drugs, although obviously related to women pregnancy, each works on different levels and also on different parts of the human body. The goal is the same, but the means to achieve it are a bit different. This kind of relation is correct but it is worth mentioning that it is also the first occurrence of non-direct association on the results regarding drug to drug co-relation. Finally, line 14 relates glipizide with testosterone. As written earlier, glipizide is used for controlling hyperglycemia in patients suffering from diabetes type II and testosterone stimulates sexual behaviors and sexual drive, and is used for managing hypogonadism and andropause. At first sight, a relation does not seem obvious but upon a brief research, we found that doctors can administer testosterone injections and follow patients' improved glycemic control, reducing their insulin requirements accordingly [41]. This is also an indirect association but plausible.

4.4.4 Co-Occurrences between Disorders and Symptoms

	Disorder	Symptom
1	Metrorrhagia	Menstrual spotting
2	Amnesia	Memory Loss
3	Pharyngitis	Sore Throat
4	Tension	Feeling tense
5	Exanthema	Spots on skin
6	Self hatred	Imbalance
7	Premature Ejaculation	Memory Loss
8	Loss of interest	Imbalance
9*	Prostate carcinoma	Blurred vision
10*	Memory impairment	Blurred vision
11	Chronic Fatigue Syndrome	Trembling
12	Erectile dysfunction	Memory Loss
13	Depersonalization	Depressive Symptoms
14	Derealization	Depressive Symptoms
15	Memory impairment	Memory Loss

Table 4.7: Top co-occurrences between disorders and symptoms.
* - denotes inconclusive relation

Similarly to the previous table, co-occurrences between disorders and symptoms also show direct associations (sometimes synonyms) but, in this case, they are almost self explanatory. Taking advantage of this fact, we will group those associations and will not expand too much on them. Associations number 2, 3, 4, 13, 14 and 15 are the ones easy to understand. For the first association, metrorrhagia is defined by being an uterine bleeding at irregular intervals, particularly between the expected menstrual periods and menstrual spotting is just another name for the same phenomenon [87, 62]. Next, we will focus on number 5: exanthema and spots on skin. Exanthema is a widespread rash on human skin normally associated with well known diseases namely measles (or rubeola) and chickenpox (or varicella) just to name a couple which are also known to leave red spots on skin [85]. As such, this association is also correct. Moving to lines 6 and 8 in which self hatred and loss of interest are related to imbalance. Unfortunately, imbalance is also a very ambiguous term and, as such, drawing conclusions can be quite difficult. Anyway, if imbalance is related to the physical lack of balance, then this association is almost impossible to explain. However, if imbalance actually means chemical imbalance, then there is an hypothesis of being the cause of those mental illnesses [55]. In this case, the second option is the one making sense and, as such, associations 6 and 8 are assumed to be

plausible. For lines 7 and 12, memory loss is associated with premature ejaculation and erectile dysfunction. Remembering what we stated earlier, sildenafil (or viagra) is commonly associated with these disorders and its side effects can cause memory loss. So, if a patient suffering from these disorders and is using sildenafil as treatment, these two associations are correct. The next two associations, 9 and 10, relate prostate carcinoma and memory impairment with blurred vision. Prostate cancer (and other forms of cancer) when in an advanced stage can metastasize to different organs and, if the cancer metastasizes to the brain, blurred vision is one of the symptoms [78]. Regarding memory impairment, there is no direct relation to blurred vision but perhaps a different condition is triggering both memory impairment and blurred vision? Lastly, row 11 relates chronic fatigue syndrome with trembling. Previously we discussed chronic fatigue syndrome as being a disorder characterized by tiredness and lack of energy and, one of the symptoms of this disorder is trembling, validating this relation [60].

4.4.5 Co-Occurrences between Disorders

	Disorder	Disorder
1	Melanocytic nevus	Benign melanocytic nevus
2*	Robinow Syndrome	Duane Retraction Syndrome
3	Child attention deficit disorder	Attention Deficit Disorder
4	Acne	Acne Vulgaris
5	Panic Disorder 1	Panic Disorder
6	Depressive disorder	Mental Depression
7	Unipolar Depression	Major Depressive Disorder
8	Dental caries	Cavitation
9	Chronic Obstructive Airway Disease	Common Cold
10	Candidiasis	Oral candidiasis
11	Cicatrization	Cicatrix
12	Communicable Diseases	Infection
13	Loss of interest	Self hatred
14	Depersonalization	Derealization
15	Self hatred	Phobic anxiety disorder

Table 4.8: Top co-occurrences between disorders.
* - denotes inconclusive relation

Analyzing co-occurrences between disorders is probably the easiest when compared to the other results presented in this chapter. These associations are related because the majority of the concepts end up being synonyms. Although not necessarily unexpected, this situation can be avoided by performing a more detailed semantic analysis for filtering associations of this kind. Associations 1, 3, 4, 5, 6, 7, 10 and 11 are perceived as spot on even for persons without medical or biological background. In any case, there are a few strange associations and, as such, we will focus on those. On line 2 we could not find a relation between one disorder and the other. Being this a work focused on finding different or unexpected concept pairings, this association is a good example of what we aimed for. Moving to line 8, this relation is considered correct as long as cavitation is assumed to be a formation of an oral cavity [26]. Dental caries are also called cavities, carious lesions or, simpler, holes in the teeth [18]. However, if cavity is assumed to be associated with cavities formed in lungs, then the diagnosis leads to a tuberculosis related disorder and, as such, invalidating the association. Line 12 is also proves to be a valid association as communicable diseases (or infectious diseases) are illnesses resulting from an infection [86]. Next associations are 13 and 15, which relate loss of interest and phobic anxiety disorder with self hatred. Self-hatred can be described as

an extreme dislike of oneself and can result from an inferiority complex; using this definition, it is plausible to associate with loss of interest and also with anxiety. In any case, mental disorders are hugely complex and this kind of relations can probably be connected but these conclusions must be drawn by whom has the expertise to do so. To end this analysis, line 14 relates depersonalization with derealization. Both disorders are characterized by an alteration in the perception of oneself (depersonalization) or the world (derealization). The association of these two disorders is also subject of discussion, with most authors currently regard both disorders as independent, others do not [83, 84].

4.5 Final Words

After the detailed analysis performed on the previous section, we found that our methods can achieve good results but also can trigger some debatable associations. Obviously, these methods are not infallible and we must find ways to improve them. One of the biggest problems we had was to find ways to clean the dictionaries in order to filter most false positives and, in a way, we achieved that. From the analysis above, it is perceptible that the results from associations between drugs/disorders and drugs/symptoms differ from the associations between disorders/symptoms, drugs/drugs and disorders/disorders. The first two cases return better results (better in a sense of more side-effects and even placebo than known relations) because we built a solution to avoid and filter known and expected associations. This solution could be extended to the other three types of relations and, eventually, yielding better results, although one of the main problems would still be the fact that most concepts identified are synonyms. Without performing a semantic analysis to identify these cases, we could use them to populate a list of previously identified synonyms and use it to filter the newer results. The problem with this solution being the fact that the identification of such synonym relations would have to be performed manually as we did in the previous chapter or by using an external agent to feed us synonyms such as the Big Huge Thesaurus API [7]. With our solution being dictionary based, one way of obtaining a broader range of results would be to expand the dictionaries scope. On the other hand, if we aim for a more specific scope or area of investigation, then the dictionaries could be cleaned and fine tuned for that specific goal. From a statistical point of view, we can always debate if our results would be different if we had used other metrics rather than Odds Ratio. It is a valid assertion but each metric also has its pros and cons and, as such, results may not have varied much.

Chapter 5

Conclusions and Closing Remarks

The main goal of this project was to identify health concepts in social networks. In order to do this, we researched and chose the sources we thought to be the adequate and after studying the available tools and frameworks, we began to implement the system. Our implementation allows expanding to different areas, for instance, given a different set of dictionaries, we can aim the scope of our research to very specific goals. Or if we want to add a new data source, we can use the structure already built and the system will work seamlessly. We had to overcome some technical issues like adapting some tools and methods to our specific needs (as described in this document) but, in the end, we are happy with our results. We proved the possibility of extracting health concepts (in the form of drugs, disorders and/or symptoms) from social networks like Yahoo Answers and Reddit. During the nine month period featured in our study, the system was able to gather over 900000 entries, which result in more than one million annotations between drug, disorder and symptom concepts. Not only were we able to extract health concepts but we also managed to find ways to co-relate drugs with disorder and symptom concepts with very satisfying and encouraging results. Furthermore, by using a slightly different analysis process, we also can understand how diseases (also their symptoms) and drugs vary over a given period of time.

For future enhancements, the addition of other sources of data should be considered. Moreover, we would like to implement the ability to store different kinds of information such as geographical data. This will enable us to pinpoint where something is happening almost in real time. Another valuable feature for this kind of work, is to broaden the language scope, or in other words, find ways to analyze different languages and not only English to further improve the range of this particular data mining system. A machine learning solution could also be considered as a further enhancement for the overall system improvement.

From a personal point of view, designing and building a solution capable of handling large amounts of data led me to search for and question different approaches and different programming languages as well as compelling me to always find ways for improving the whole system. Performing analysis in large amounts of data is a very demanding and time consuming task, and this was probably the part of the project which took me more time to accomplish: trying to find ways of speeding up the analysis process to an acceptable level. The most non-technical challenge of this project was, perhaps, to work with information that I don't necessarily have the expertise to analyze. On the other hand, it is also one of the things that makes this sort of projects so compelling. In the end, and after working for some months on this subject, I learnt why more and more companies rely on data mining systems to further understand what common people (clients or others) do and expect from a determined product or service.

Appendix

Appendix A - Disorder and Drug Classes

Disorder and symptom classes extracted from the Comparative Toxicogenomics Database

Classes	
Animal disease	Metabolic disease
Bacterial infection or mycosis	Mouth disease
Blood disease	Musculoskeletal disease
Cancer	Nervous system disease
Cardiovascular disease	Nutrition disorder
Congenital abnormality	Occupational disease
Connective tissue disease	Parasitic disease
Digestive system disease	Pathology (anatomical condition)
Ear-nose-throat disease	Pathology (process)
Endocrine system disease	Pregnancy complication
Environmental origin disorders	Respiratory tract disease
Eye disease	Signs and symptoms
Fetal disease	Skin disease
Genetic disease (inborn)	Substance-related disorder
Immune system disease	Urogenital disease (female)
Infant-newborn disease	Urogenital disease (male)
Lymphatic disease	Viral disease
Mental disorder	Wounds and injuries

Table 5.1: List of disorder and symptom classes.

Drug classes extracted from DrugBank

Classes	
Alkaloids and alkaloid derivatives	Nitroimidazoles
Amino acids	Phenethylamines
Benzene and derivatives	Phenols and derivatives
Benzodiazepines	Phenylacetates
Benzofurans	Phenylpropylamines
Biguanides	Pterins
Diphenylmethanes	Steroids and steroid derivatives
Inorganic ions and gases	Tametralines
Morphinans	Tocopherols
Morphine and derivatives	Xanthines

Table 5.2: List of drug classes.

Disorder and drug classes and respective number of individual occurrences

	Drug classes		Disorder classes	
1	steroids and steroid derivatives	1871	Signs and symptoms	187205
2	amino acids	1619	Mental disorder	151881
3	alkaloids and alkaloid derivatives	1423	Nervous system disease	120969
4	xanthines	1202	Respiratory tract disease	55328
5	phenylacetates	1038	Skin disease	50656
6	phenols and derivatives	719	Pathology (process)	50166
7	inorganic ions and gases	612	Cancer	37857
8	phenylpropylamines	440	Viral disease	36943
9	benzofurans	420	Bacterial infection or mycosis	35708
10	benzodiazepines	397	Musculoskeletal disease	28485
11	phenethylamines	385	Urogenital disease (female)	24874
12	morphinans	384	Genetic disease (inborn)	21026
13	morphine and derivatives	379	Substance-related disorder	19329
14	diphenylmethanes	309	Immune system disease	19288
15	benzene and derivatives	260	Mouth disease	17946
16	biguanides	257	Urogenital disease (male)	17736
17	tametralines	225	Cardiovascular disease	17037
18	pterins	202	Metabolic disease	16941
19	tocopherols	162	Digestive system disease	16606
20	nitroimidazoles	148	Endocrine system disease	15253

Table 5.3: List of drug and disorder classes and respective occurrences.

Appendix B - Lists of known concept co-occurrences

Known co-occurrences between drugs and disorders

	Drug	Disorder
1	cyanocob(III)alamin	Anemia
2	metformin	Polycystic Ovary Syndrome
3	metformin	Insulin Resistance
4	cortancyl	Crohn Disease
5	L-thyroxine	Hypothyroidism
6	clomiphene	Polycystic Ovary Syndrome
7	testosterone	Erectile dysfunction
8	thiamine(1+)	Autistic Disorder
9	calciol	Autistic Disorder
10	tramadol	Fibromyalgia
11	progesterone	Infertility
12	metformin	Diabetes Mellitus, Non-Insulin-Dependent
13	sertraline	Paranoia
14	clomiphene	Spontaneous abortion
15	diphenhydramine	Urticaria
16	progesterone	Polycystic Ovary Syndrome
17	lamotrigine	Bipolar Disorder
18	diazepam	Panic Disorder
19	progesterone	Spontaneous abortion
20	cholesterol	Heart Diseases
21	miconazole	Candidiasis
22	metformin	Endometriosis, site unspecified
23	lamotrigine	Mood Disorders
24	potassium chloride	leukemia
25	fructose	Insulin Resistance
26	lithium tetrahydroaluminate	Bipolar Disorder
27	famotidine	Crohn Disease
28	miconazole	Oral candidiasis
29	metformin	Diabetes Mellitus
30	metformin	Hypoglycemia
31	cortancyl	Inflammation
32	famotidine	Irritable Bowel Syndrome
33	sertraline	Obsessive-Compulsive Disorder
34	potassium chloride	Malignant tumor of colon
35	famotidine	Celiac Disease

Table 5.4: List of known drug and disorder co-occurrences.

Known co-occurrences between drug and symptoms

	Drug	Symptom
1	cortancyl	Arthralgia
2	venlafaxine	Xerostomia
3	methylphenidate	Xerostomia
4	famotidine	Dyspepsia
5	melatonin	Asthenia
6	progesterone	Hot flushes
7	cyanocob(III)alamin	Dizziness
8	potassium	Spasm
9	famotidine	Heartburn
10	bupropion	Xerostomia
11	(S)-duloxetine hydrochloride	Xerostomia
12	morphine	Abdominal Pain
13	melatonin	Fatigue
14	aripiprazole	Agitation
15	quetiapine	Agitation
16	cyanocob(III)alamin	Headache
17	norethisterone	Dyspepsia
18	cyanocob(III)alamin	Nausea
19	morphine	Spasm
20	famotidine	Abdominal Pain
21	fluoxetine	Hot flushes
22	venlafaxine	Agitation
23	diphenhydramine	Pruritus
24	bupropion	Agitation
25	progesterone	Mastodynia
26	diazepam	Chest Pain
27	lithium tetrahydroaluminate	Xerostomia
28	lithium tetrahydroaluminate	Agitation
29	(R)-adrenaline	Dyspnea
30	tramadol	Back Pain
31	methylphenidate	Agitation
32	magnesium(2+)	Constipation
33	paroxetine	Xerostomia
34	sertraline	Headache
35	cortancyl	Diarrhea

Table 5.5: List of known drug and symptom co-occurrences.

Appendix C - Lists of concept co-occurrences

Co-occurrences between drugs and disorders

	Drug	Disorder
1	caffeine	caffeine stimulant related disorder
2	glipizide	Prostate carcinoma
3	glipizide	Malignant neoplasm of prostate
4	miconazole	Yeast infection
5	glipizide	Memory impairment
6	sertraline	Self hatred
7	cyanocob(III)alamin	Chronic Fatigue Syndrome
8	infliximab	Crohn Disease
9	sertraline	Loss of interest
10	sildenafil	Premature Ejaculation
11	infliximab	Cancer Remission
12	adalimumab	Crohn Disease
13	glipizide	Psoriasis
14	glipizide	insanity
15	clomiphene	Infertility
16	cyanocob(III)alamin	Infectious Mononucleosis
17	thiamine(1+)	Autism Spectrum Disorders
18	calciol	Autism Spectrum Disorders
19	adalimumab	Cancer Remission
20	glipizide	Blind Vision
21	testosterone	Prostate carcinoma
22	testosterone	Premature Ejaculation
23	sildenafil	Amnesia
24	testosterone	Malignant neoplasm of prostate
25	sertraline	Hypochondriasis
26	fructose	Corn of toe
27	miconazole	Bacterial Infections
28	thiamine(1+)	Atrial Septal Defects
29	calciol	Atrial Septal Defects
30	morphine	Opioid abuse
31	(S)-duloxetine hydrochloride	Fibromyalgia
32	glipizide	Amnesia
33	cyanocob(III)alamin	Post-Traumatic Stress Disorder
34	oxycodone	Opioid abuse
35	cortancyl	Cancer Remission

Table 5.6: List of drug and disorder co-occurrences.

Co-occurrences between drugs and symptoms

	Drug	Symptom
1	cyanocob(III)alamin	Trembling
2	sildenafil	Memory Loss
3	glipizide	Memory Loss
4	infliximab	Flare
5	cyanocob(III)alamin	Weakness
6	glipizide	Back Pain
7	melatonin	Trembling
8	adalimumab	Flare
9	cortancyl	Flare
10	cyanocob(III)alamin	Lassitude
11	glipizide	Sleeplessness
12	glipizide	Stomach ache
13	cyanocob(III)alamin	Asthenia
14	testosterone	Blurred vision
15	glipizide	Ache
16	sertraline	Imbalance
17	adalimumab	Arthralgia
18	infliximab	Arthralgia
19	cyanocob(III)alamin	Sensory Discomfort
20	cyanocob(III)alamin	Fatigue
21	cyanocob(III)alamin	Lightheadedness
22	clomiphene	Hot flushes
23	clomiphene	Moaning
24	cyanocob(III)alamin	Vertigo
25	diazepam	Hot flushes
26	(S)-nicotine	Withdrawal Symptoms
27	chloramphenicol	Knee pain
28	phenylephrine	Nasal congestion (finding)
29	testosterone	Memory Loss
30	sildenafil	Blurred vision
31	melatonin	Weakness
32	lisdexamfetamine	Xerostomia
33	melatonin	Sleeplessness
34	dextroamphetamine	Xerostomia
35	miconazole	Abdominal bloating

Table 5.7: List of drug and symptom co-occurrences.

Co-occurrences between drugs

	Drug	Drug
1	thiamine(1+)	calciol
2	infliximab	adalimumab
3	cortancyl	infliximab
4	amphetamine	dextroamphetamine
5	lamotrigine	lithium tetrahydroaluminate
6	adalimumab	cortancyl
7	lamotrigine	aripiprazole
8	lisdexamfetamine	dextroamphetamine
9	aripiprazole	lithium tetrahydroaluminate
10	lisdexamfetamine	amphetamine
11	oxycodone	morphine
12	clomiphene	progesterone
13	quetiapine	aripiprazole
14	glipizide	testosterone
15	quetiapine	lamotrigine
16	zinc atom	calcium
17	amphetamine	methylphenidate
18	lisdexamfetamine	methylphenidate
19	dextroamphetamine	methylphenidate
20	lithium tetrahydroaluminate	valproic acid
21	clomiphene	metformin
22	tramadol	(S)-duloxetine hydrochloride
23	magnesium(2+)	potassium
24	oxycodone	tramadol
25	lamotrigine	valproic acid
26	valproic acid	quetiapine
27	11-cis-retinol	(+)-alpha-tocopherol
28	aripiprazole	valproic acid
29	cocaine	3,4-methylenedioxymethamphetamine
30	acetaminophen	oxycodone
31	(S)-duloxetine hydrochloride	amitriptyline
32	zinc atom	calciol
33	clonazepam	alprazolam
34	(S)-duloxetine hydrochloride	venlafaxine
35	bupropion	aripiprazole

Table 5.8: List of drug co-occurrences.

Co-occurrences between disorders and symptoms

	Disorder	Symptom
1	Metrorrhagia	Menstrual spotting
2	Amnesia	Memory Loss
3	Pharyngitis	Sore Throat
4	Tension	Feeling tense
5	Exanthema	Spots on skin
6	Self hatred	Imbalance
7	Premature Ejaculation	Memory Loss
8	Loss of interest	Imbalance
9	Prostate carcinoma	Blurred vision
10	Memory impairment	Blurred vision
11	Chronic Fatigue Syndrome	Trembling
12	Erectile dysfunction	Memory Loss
13	Depersonalization	Depressive Symptoms
14	Derealization	Depressive Symptoms
15	Memory impairment	Memory Loss
16	Hypochondriasis	Imbalance
17	Malignant neoplasm of prostate	Memory Loss
18	Hypochondriasis	Depressive Symptoms
19	Schizophrenia	Verbal auditory hallucinations
20	Anemia	Trembling
21	Asthma	Wheezing
22	Gastroesophageal reflux disease	Heartburn
23	Crohn Disease	Flare
24	Premature Ejaculation	Back Pain
25	Infection of ear	Earache
26	Prostate carcinoma	Back Pain
27	Phobic anxiety disorder	Imbalance
28	Malignant neoplasm of prostate	Back Pain
29	Self hatred	Headache
30	Chronic Fatigue Syndrome	Fatigue
31	Chronic Fatigue Syndrome	Weakness
32	Lymphoma	Wheezing
33	Psoriasis	Blurred vision
34	Gastroesophageal reflux disease	Heartburn acidity
35	Malignant neoplasm of skin	Wheezing

Table 5.9: List of disorder and symptom co-occurrences.

Co-occurrences between disorders

	Disorder	Disorder
1	Melanocytic nevus	Benign melanocytic nevus
2	Robinow Syndrome	Duane Retraction Syndrome
3	Child attention deficit disorder	Attention Deficit Disorder
4	Acne	Acne Vulgaris
5	Panic Disorder 1	Panic Disorder
6	Depressive disorder	Mental Depression
7	Unipolar Depression	Major Depressive Disorder
8	Dental caries	Cavitation
9	Chronic Obstructive Airway Disease	Common Cold
10	Candidiasis	Oral candidiasis
11	Cicatrization	Cicatrix
12	Communicable Diseases	Infection
13	Loss of interest	Self hatred
14	Self hatred	Hypochondriasis
15	Depersonalization	Derealization
16	Self hatred	Phobic anxiety disorder
17	Erectile dysfunction	Premature Ejaculation
18	Loss of interest	Hypochondriasis
19	Self hatred	Obsessions
20	Prostate carcinoma	Premature Ejaculation
21	Malignant neoplasm of prostate	Premature Ejaculation
22	Self hatred	Paranoia
23	Bronchopulmonary Dysplasia	Borderline Personality Disorder
24	Memory impairment	Premature Ejaculation
25	Memory impairment	Prostate carcinoma
26	Memory impairment	Malignant neoplasm of prostate
27	Loss of interest	Phobic anxiety disorder
28	Loss of interest	Obsessions
29	Self hatred	Aggressive behavior
30	Prostate carcinoma	Erectile dysfunction
31	Malignant neoplasm of prostate	Erectile dysfunction
32	Loss of interest	Paranoia
33	Herpes NOS	Genital Herpes
34	Self hatred	Eating Disorders
35	Self hatred	Osteochondritis Dissecans

Table 5.10: List of disorder co-occurrences.

Bibliography

[1] 1uponcancer.com - LEARN MORE ABOUT YOUR DRUGS THAN IS ON THE LABEL. http://www.1uponcancer.com/2013/09/27/learn-more-about-your-drugs-than-is-on-the-label/#more-7204, 2014.

[2] A live look at activity across WordPress.com. http://en.wordpress.com/stats/, 2014.

[3] Phenyl acetate - Wikipedia. http://en.wikipedia.org/wiki/Phenyl_acetate, 2014.

[4] Yahoo Answers: Compete Site Analytics. https://siteanalytics.compete.com/answers.yahoo.com/, 2013.

[5] GREC Corpus - Standoff annotation format. http://www.biomedcentral.com/content/supplementary/1471-2105-10-349-s1/standoff.html, 2010.

[6] Yahoo Answers. http://answers.yahoo.com, 2014.

[7] Big Huge Thesaurus API. http://words.bighugelabs.com/api.php, 2014.

[8] Reddit API. https://github.com/reddit/reddit/wiki/API, 2014.

[9] Yahoo Answers API. https://developer.yahoo.com/answers/, 2014.

[10] Java ArrayList. http://docs.oracle.com/javase/7/docs/api/java/util/ArrayList.html, 2014.

[11] Lars Backstrom and Jon M. Kleinberg. *Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook.* CoRR, volume abs/1310.6753, 2013.

[12] Fawzi Elias Bekri and Dr. A. Govardhan. *Association of Data Mining and healthcare domain: Issues and current state of the art*, 2011.

[13] Stephen J Bickston. *Infliximab for Ulcerative Colitis Induction of Remission and Maintenance Therapy.* 2007.

[14] J Martin Bland and Douglas G Altman. *The odds ratio.* BMJ, volume 320, no. 7247, p. 1468, 2000. doi:10.1136/bmj.320.7247.1468.

[15] K.Paul Bouter, Rob J.A. Diepersloot, Leo K.J. van Romunde, Roeland Uitslager, Nic Masurel, Joost B.L. Hoekstra and D. Willem Erkelens. *Effect of epidemic influenza on ketoacidosis, pneumonia and death in diabetes mellitus: a hospital register survey of 1976-1979 in The Netherlands.* Diabetes Research and Clinical Practice, volume 12, no. 1, pp. 61 – 68, 1991. doi:http://dx.doi.org/10.1016/0168-8227(91)90131-V.

[16] Twitter Statistics: Statistic Brain. `http://www.statisticbrain.com/twitter-statistics/`, 2014.

[17] David Campos, Sergio Matos and Jose Oliveira. *A modular framework for biomedical concept recognition.* BMC Bioinformatics, volume 14, no. 1, p. 281, 2013. doi:10.1186/1471-2105-14-281.

[18] Dental caries - Wikipedia. `http://en.wikipedia.org/wiki/Dental_caries`, 2014.

[19] Ae Chun, Soon, MacKellar and Bonnie. *Social Health Data Integration Using Semantic Web.* In Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, pp. 392–397. ACM, New York, NY, USA, 2012. doi:10.1145/2245276.2245351.

[20] cron4j - a scheduler for the Java platform. `http://www.sauronsoftware.it/projects/cron4j/index.php`, 2012.

[21] crontab - schedule periodic background work. `http://pubs.opengroup.org/onlinepubs/9699919799/utilities/crontab.html`, 2013.

[22] BNC database and word frequency lists. `http://www.kilgarriff.co.uk/bnc-readme.html`, 2006.

[23] CTD - Comparative Toxicogenomics Database. `http://ctdbase.org/`, 2014.

[24] MedicineNet.com - Xanthine derivatives oral. `http://www.medicinenet.com/xanthine_derivatives-oral/article.htm`, 2014.

[25] The Free Dictionary: Medical dictionary - definition of flare. `http://medical-dictionary.thefreedictionary.com/flare`, 2014.

[26] The Free Dictionary: Medical dictionary - definition of flare. `http://medical-dictionary.thefreedictionary.com/cavitation`, 2014.

[27] Comparative Toxicogenomics Database - Chemical disease associations. `http://ctdbase.org/downloads/#cd`, 2014.

[28] Son Doan, Ai Kawazoe and Nigel Collier. *N.: Global Health Monitor - a web-based system for detecting and mapping infectious diseases.* In In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). (2008) 951-956. 2008.

[29] eHealthMe.com - FDA and Personalized Social Media. `http://www.ehealthme.com/`, 2014.

[30] eHealthMe.com - Review: could Humira cause Rashes? `http://www.ehealthme.com/ds/Humira/Rashes`, 2014.

[31] eHealthMe.com - Review: could Remicade cause Rashes? `http://www.ehealthme.com/ds/remicade/Rashes`, 2014.

[32] eHealthMe.com - Review: Glipizide side effects. `http://www.ehealthme.com/glipizide/glipizide-side-effects`, 2014.

[33] G. Eysenbach. *Infodemiology: tracking flu-related searches on the web for syndromic surveillance.* In Proceedings of AMIA 2006 Annual Symposium, pp. 244–248. 2006.

[34] Facebook. `http://www.facebook.com`, 2014.

[35] Facebook Key Facts. `http://newsroom.fb.com/Key-Facts`, 2013.

[36] families.com - Why Kids with Autism May Need Vitamin Supplements. `http://www.families.com/blog/why-kids-with-autism-need-vitamin-supplements`, 2014.

[37] JReddit - A Java API for accessing Reddit. `https://github.com/acbart/JReddit`, 2013.

[38] Sherine E. Gabriel, Liisa Jaakkimainen and Claire Bombardier. *Risk for Serious Gastrointestinal Complications Related to Use of Nonsteroidal Anti-inflammatory DrugsA Meta-analysis.* Annals of Internal Medicine, volume 115, no. 10, pp. 787–796, 1991. doi: 10.7326/0003-4819-115-10-787.

[39] Frank D Gianfrancesco, Amy L Grogg, Ramy A Mahmoud, Ruey-hua Wang and Henry A Nasrallah. *Differential effects of risperidone, olanzapine, clozapine, and conventional antipsychotics on type 2 diabetes: findings from a large health plan database.* The Journal of clinical psychiatry, volume 63, no. 10, pp. 920–930, 2002.

[40] google-gson - A Java library to convert JSON to Java objects and vice-versa. `http://code.google.com/p/google-gson/`, 2013.

[41] Life Extension Magazine - Testosteroneś Overlooked Role in the Treatment of Diabetes in Men. `http://www.lef.org/magazine/2007/7/report_diabetes/page-01`, 2007.

[42] JSON - JavaScript Object Notation. http://www.json.org/, 2009.

[43] Jaeseung Jeong, John C Gore and Bradley S Peterson. *Mutual information analysis of the {EEG} in patients with Alzheimer's disease.* Clinical Neurophysiology, volume 112, no. 5, pp. 827 – 835, 2001. doi:http://dx.doi.org/10.1016/S1388-2457(01)00513-2.

[44] Vasileios Lampos, Tijl De Bie and Nello Cristianini. *Flu Detector - Tracking Epidemics on Twitter.* In JoséLuis Balcázar, Francesco Bonchi, Aristides Gionis and Michèle Sebag, editors, Machine Learning and Knowledge Discovery in Databases, volume 6323 of *Lecture Notes in Computer Science*, pp. 599–602. Springer Berlin Heidelberg, 2010. doi:10.1007/978-3-642-15939-8_42.

[45] Yahoo Query Language. https://developer.yahoo.com/yql/, 2014.

[46] V. Law, C. Knox, Y. Djoumbou, T. Jewison, AC. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, ZT. Dame, B. Han, Y. Zhou and DS. Wishart. *DrugBank 4.0: shedding new light on drug metabolism.* volume 42, no. 1, pp. D1091–7, 2014.

[47] Livestrong - Negative Side Effects of Testosterone. http://www.livestrong.com/article/31320-negative-side-effects-testosterone/, 2011.

[48] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal and P. Suetens. *Multimodality image registration by maximization of mutual information.* Medical Imaging, IEEE Transactions on, volume 16, no. 2, pp. 187–198, 1997. doi:10.1109/42.563664.

[49] Bliss GVS - Pessary / Vaginal Suppository Manufacturer and Exporter. http://www.blissgvs.com/products/pharma-products/pessaries, 2014.

[50] Abla Mawudeku, Michael Blench, Louise Boily, Ron St. John, Roberta Andraghetti and Martha Ruben. The Global Public Health Intelligence Network, pp. 457–469. John Wiley and Sons Ltd, 2013. doi:10.1002/9781118543504.ch37.

[51] CG McMahon, BG Stuckey, M Andersen, K Purvis, N Koppiker, S Haughie and M Boolell. *Efficacy of sildenafil citrate (Viagra) in men with premature ejaculation.* 2005.

[52] EC McNaughton, PM Coplan, RA Black, SE Weber, HD Chilcoat and SF Butler. *Monitoring of internet forums to evaluate reactions to the introduction of reformulated OxyContin to deter abuse.* 2014. doi:10.2196/jmir.3397.

[53] PatientsLikeMe Newsroom: Disease Milestones. http://news.patientslikeme.com/milestones, 2014.

[54] Stephen S. Morse, Barbara Hatch Rosenberg and Jack Woodall. *ProMED global monitoring of emerging diseases: design for a demonstration program.* Health Policy, volume 38, no. 3, pp. 135 – 153, 1996. doi:http://dx.doi.org/10.1016/0168-8510(96)00863-9.

[55] mymentalhealth.ca - Learn About Mental Illness. `http://www.mymentalhealth.ca/learn/causes/tabid/839/default.aspx`, 2014.

[56] H Nakada, K Yuji, M Tsubokura, Y Ohsawa and M Kami. *Development of a national agreement on human papillomavirus vaccination in Japan: an infodemiology study.* 2014. doi:10.2196/jmir.2846.

[57] Neji. `http://bioinformatics.ua.pt/neji/`, 2014.

[58] Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick and Margaret E. Morris. *It's Not That I Don'T Have Problems, I'M Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health.* In Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11, pp. 341–350. ACM, New York, NY, USA, 2011. doi:10.1145/1958824.1958876.

[59] List of existing Reddit API Wrappers - Reddit. `http://www.reddit.com/r/redditdev/comments/nd521/list_of_existing_reddit_api_wrappers/`, 2014.

[60] University of Maryland Medical Center - Chronic fatigue syndrome. `http://umm.edu/health/medical/altmed/condition/chronic-fatigue-syndrome`, 2012.

[61] PatientsLikeMe. `http://www.patientslikeme.com`, 2014.

[62] Healthline - Vaginal Bleeding Between Periods. `http://www.healthline.com/health/vaginal-bleeding-between-periods#Overview1`, 2014.

[63] Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson and Robert A. Weinstein. *Using Internet Searches for Influenza Surveillance.* Clinical Infectious Diseases, volume 47, no. 11, pp. 1443–1448, 2008. doi:10.1086/593098.

[64] PostgreSQL - Wikipedia. `http://en.wikipedia.org/wiki/PostgreSQL`, 2014.

[65] PostgreSQL Global Development Group. `http://www.postgresql.org`, 2014.

[66] Víctor M. Prieto, Sergio Matos, Manuel Álvarez, Fidel Cacheda and José Luís Oliveira. *Analysing Relevant Diseases from Iberian Tweets.* In Mohd Saberi Mohamad, Loris Nanni, Miguel P. Rocha and Florentino Fdez-Riverola, editors, 7th International Conference on Practical Applications of Computational Biology & Bioinformatics, volume 222 of *Advances in Intelligent Systems and Computing*, pp. 69–76. Springer International Publishing, 2013. doi:10.1007/978-3-319-00578-2_10.

[67] PubMed.gov. http://www.ncbi.nlm.nih.gov/pubmed/, 2014.

[68] He Qian, E. Agu, D. Strong, B. Tulu and P. Pedersen. *Characterizing the Performance and Behaviors of Runners Using Twitter.* In Healthcare Informatics (ICHI), 2013 IEEE International Conference on, pp. 406–414. 2013. doi:10.1109/ICHI.2013.56.

[69] Schulz R and Beach SR. *Caregiving as a risk factor for mortality: The caregiver health effects study.* JAMA, volume 282, no. 23, pp. 2215–2219, 1999. doi:10.1001/jama.282.23.2215.

[70] Sean A Savitz and Louis R Caplan. *Transient global amnesia after sildenafil (Viagra) use.* Neurology, volume 59, no. 5, p. 778, 2002. doi:10.1212/WNL.59.5.778.

[71] AYTM - Private Profiles Survey: Most Social Media Users Apply Privacy Settings. https://aytm.com/blog/daily-survey-results/private-profiles-survey/, 2013.

[72] Alessio Signorini, Alberto Maria Segre and Philip M. Polgreen. *The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic.* PLoS ONE, volume 6, no. 5, p. e19467, 2011. doi:10.1371/journal.pone.0019467.

[73] Reddit Traffic Statistics. http://www.reddit.com/r/AskReddit/about/traffic, 2014.

[74] Teens Health - Are steroids worth the risk? http://kidshealth.org/teen/food_fitness/sports/steroids.html, 2014.

[75] How stuff works - Alkaloid. http://science.howstuffworks.com/alkaloid-info.htm, 2014.

[76] UMLS - Unified Medical Language System. http://www.nlm.nih.gov/research/umls/, 2014.

[77] Simply Singleton - Navigate the deceptively simple Singleton pattern. http://www.javaworld.com/article/2073352/core-java/simply-singleton.html, 2003.

[78] Cancer Research UK - Brain tumour symptoms. http://www.cancerresearchuk.org/about-cancer/type/brain-tumour/about/brain-tumour-symptoms, 2014.

[79] Twitter. http://www.twitter.com, 2014.

[80] Treato - The voice of the patient. http://www.treato.com/, 2014.

[81] WebMD. http://www.webmd.com/default.htm, 2014.

[82] Viagra Labels to Note Amnesia WebMD - Cialis. http://www.webmd.com/erectile-dysfunction/news/20080822/cialis-viagra-labels-to-note-amnesia, 2008.

[83] Depersonalization - Wikipedia. http://en.wikipedia.org/wiki/Depersonalization, 2014.

[84] Derealization - Wikipedia. http://en.wikipedia.org/wiki/Derealization, 2014.

[85] Exanthema - Wikipedia. http://en.wikipedia.org/wiki/Exanthem, 2014.

[86] Infection - Wikipedia. http://en.wikipedia.org/wiki/Infection, 2014.

[87] Metrorrhagia - Wikipedia. http://en.wikipedia.org/wiki/Metrorrhagia, 2014.

[88] Testosterone - Wikipedia. http://en.wikipedia.org/wiki/Testosterone, 2014.

[89] E Yom-Tov, D Borsa, IJ Cox and RA McKendry. *Detecting disease outbreaks in mass gatherings using internet data.* 2014. doi:10.2196/jmir.3156.